SEQUENCE ANALYSIS OF THE I FACTOR FROM *DROSOPHILA MELANOGASTER*

DIANA H. FAWCETT

Ph.D. THESIS

EDINBURGH UNIVERSITY

1987

EX BIBL. UNIV. EDINBURGEN.

Acknowledgements

I would like to thank my supervisor, Dr David Finnegan, for supervising this project and for his help and advice during the last four years.

I would also like to thank Heather Houston and Dr Arthur Robinson for many helpful discussions concerning DNA sequencing, Dr Noreen Murray for advice on lambda cloning techniques and Dr John Collins, Andrew Lyall and Dr Andrew Coulson for their invaluable aid with the computing.

My thanks go also to Jackie Bogie who produced this typescript and to Graham Brown for the photography.

Finally, I wish to thank my fiance, Mark, for his support and encouragement throughout the writing of this thesis.

ABSTRACT

This thesis is concerned with the analysis of the I factor transposable element in Drosophila melanogaster. One such element, from the strain w^{IR1}, had previously been shown to be capable of inducing hybrid dysgenesis and had been cloned. The nucleotide sequence of this element is presented in this thesis, plus an analysis of the sequence to assess coding potential and to propose a possible transposition mechanism. In addition the ends of several other I factors, cloned from D. melanogaster strains w^{IR2-6} and bx^{F31}, have been sequenced.

The I factor is 5.4kb long. Each element studied is flanked by a target site duplication, a feature characteristic of transposable element insertions. The size of the duplication varies between 9 and 14bp.

The I factor has no direct or inverted terminal repeats. Instead at the designated "right hand" end there is a short run of TAA triplets. Elements with 4, 5, 6 and 7 triplets have been found. This lack of terminal repeats likens the I factor more to the retroposon class than to any other family of transposable element in Drosophila such as the copia-like, FB and P elements.

Computer analysis of the I factor sequence has revealed two long open reading frames, both in the top strand. ORF1 is 1287bp long and ORF2 is 3258bp long. The sequence of each reading frame has been

translated and the amino acid sequences analysed. Within the ORF2 peptide, homology is found in seven domains characteristic of reverse transcriptases. It is proposed in this thesis that the I factor codes for a reverse transcriptase, which could function as a transposase by copying into DNA an RNA transcribed from the full length of the element. Similar homology is found within one of the two open reading frames of mammalian LINE sequences (members of the retroposon class of elements) and striking homology is seen between the amino acid sequences of LINEs and the I factor in this region. It is suggested that the I factor could represent the *Drosophila* equivalent of a LINE sequence.

The amino acid sequence of ORF1 of the I factor contains homology to a retroviral protein domain thought to interact with nucleic acids. The possibility that the product of ORF1 could act as a repressor of I factor activity in I strains by binding to I factor DNA or RNA is discussed.

Two additional strains, $\underline{w}^{IR7}$ and $\underline{w}^{IR8}$, were also studied. These strains carry mutations of the *white* gene and were detected following an IR dysgenic cross. These mutations are not associated with I factor insertions, instead they are deletions of part of the *white* gene. Sequence analysis of the deletion breakpoints showed no evidence of I factor activity. It is possible that the deletions arose spontaneously rather than as the result of I factor transposition.

CONTENTS

Abbreviations

bp	=	base pair
kb	=	kilobase
dNTP	=	deoxynucleotide triphosphate (e.g. dCTP, dATP)
dGAT or dGATC	=	mixture of indicated dNTPs
MOI	=	multiplicity of infection
X-gal	=	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
MTG	=	methyl-β-D-thiogalactoside
EDTA	=	ethylenediaminetetraacetic acid
SDS	=	sodium dodecyl sulphate
GD	=	gonadal dysgenesis
SF	=	stérilité femelle
LTR	=	long terminal repeat

CHAPTER 1

Introduction

1.1 Introductory comments

Transposable elements have been found in the genomic DNA of every species analysed so far. In many cases the tranposability of an element is inferred, rather than observed, either from the position of the element (e.g. within a gene), from heterogeneity of genomic location within a species, or from the structure of the element itself, and its location. Every element found so far is flanked by a direct repeat of the target site DNA sequence, thought to be generated as a result of integration. Hence this is usually regarded as an indication of transposability.

Of all species, Drosophila melanogaster has been studied most intensively with regard to transposable elements, and many different element families have been identified. For review see Finnegan and Fawcett (1986). These elements are all moderately repetitive sequences and comprise about 10% of the genome (Young, 1979). The I and P element families are amongst the most intriguing because of their involvement in the production of hybrid dysgenesis (see section 1.2). Although this phenomenon was noticed some time ago (Green, 1976) it was only in the last few years that transposable elements were recognised as being the causative agents.

1.2 Hybrid Dysgenesis

Hybrid dysgenesis is the production of abnormal characteristics in the hybrid progeny produced when certain interacting strains of Drosophila melanogaster are crossed in a particular way (Kidwell & Kidwell,

1976). Two independent systems exist, called PM and IR. In the PM system, a P (paternal) strain male must be crossed with an M (maternal) strain female for dysgensis to occur (Kidwell et al., 1977). In the IR system an I (inducer) strain male must be crossed with an R (reactive) strain female (Bucheton et al., 1976). So-called "neutral" strains exist in each system, designated Q in the PM system and N in the IR system. These have now been shown to be extremely weak P and R strains respectively, and not truly neutral (Bregliano & Kidwell, 1983). This interaction has been shown to be due to the presence, in I and P strains, of autonomous transposable elements lacking from R and M strains. Functional elements are known as I factors and P factors in the IR and PM systems respectively, defective elements are known as I and P elements. ("Element" is also frequently used as a collective term for defective and non-defective elements).

The abnormal traits produced by the two systems show many similarities. These traits include partial or complete sterility, increased mutation frequencies, chromosome rearrangements and chromosome loss. Two major differences exist however which distinguish between the two systems. Firstly, IR dysgenesis affects only hybrid female progeny (Picard et al., 1978; Proust & Prudhommeau, 1982), whereas in the PM system both sexes are affected, (although there is some evidence for the affects being more severe in females) (Engels & Preston, 1979; Kidwell & Novy, 1979). Secondly, there is the nature of the induced sterility. PM dysgenic flies exhibit a failure of the gonads of both males and females to develop properly. This is called

GD sterility (for gonadal dysgenesis). In severe cases the fly is totally sterile; in less severe cases only one side of the fly may be affected. IR dysgenesis reduces the fertility of only female progeny - these are known as SF females (for sterilité femelle). The gonads develop normally and the flies lay a normal number of eggs, but a proportion of the embryos die before reaching the blastoderm stage. Embryonic development is arrested around the 3rd or 4th nuclear division, very occasionally as late as the 6th or 7th division (Lavige, 1986). Spindle abnormalities and chromosome fragmentation cause mitosis to break down (Picard et al., 1977; Lavige and Lecher, 1982). In addition, a proportion of the embryos which pass this stage fail to hatch, possibly due to mutations and chromosome aberrations occurring in the germ cells of the SF females (Lavige, 1986). Females from the reciprocal cross (i.e. I females x R males) are called RSF females and are fully fertile.

Dysgenic events in both systems are confined to the germ line of hybrid flies - somatic mutations are very rarely seen. These first generation progeny will therefore exhibit reduced fertility but mutation events will only be seen in the soma of the second generation.

All Drosophila melanogaster strains can be classified with respect to both systems. Kidwell (1979) has grouped all strains into five categories - IP, IM, IQ, RM and NM (IQ and NM can now be included in the IP and RM groups respectively). No flies of type RP have been found, but this does not seem to be due to any biochemical constraint

as synthetic RP strains have been made and appear to be stable (Kidwell, unpublished). This pattern may reflect a sequential evolution of the I and P types. Kidwell et al. (1977) found that all long-established laboratory stocks were of the M type whereas natural populations were usually P type. Picard et al. (1976) found that all natural populations are of the I type whereas laboratory stocks could be either I, R or Q. Kidwell (1982) has proposed a "recent invasion" hypothesis to explain this distribution. Laboratory stocks collected about 50 years ago were the first to be of I type. The more recently strains were collected, the more frequent the occurrence of the I type. The occurrence of P strains follows a similar trend but do not appear in laboratory stocks collected pre-1950s. Kidwell proposes that I factors invaded natural populations about 50 years ago and P factors about 20 years later, and that these elements spread rapidly. Isolated laboratory stocks of the R and M types therefore represent relics of pre-invasion strains.

An alternative hypothesis has been proposed (Bucheton et al., 1976; Engels, 1981) whereby I and P factors have always been present in natural populations and have been lost from laboratory stocks, perhaps due to changes in environment or abnormal population structure. This is called the "stochastic loss" hypothesis.

Studies on the distribution of P and I element related sequences in other Drosophila species suggest that the spread of these elements has occurred via different mechanisms. Brookfield et al. (1984) have demonstrated that P element sequences are not found in D.

melanogaster sibling species. They are found in more distantly related species however (Daniels et al., 1984; Lansman et al., 1985), which implies that P element sequences arose in D. melanogaster as a result of some horizontal transmission event rather than vertical transmission from an evolutionary common ancestor. M strains, some of which contain no P element sequences at all, would represent strains which had not been contaminated with P sequences. There is a class of M strains which do have P-homologous sequences but behave as M strains. These are called M' strains and are believed to contain only defective P elements. These could easily be explained by loss of functional P factors, leaving only defective P elements in the genome. It seems likely then that both invasion and loss could have occurred: invasion first of all of D. melanogaster by P factors from a distantly related species, leaving M strains with no P sequences as relics of pre-invasion strains, and secondly loss from some P strains of functional P factors, to form M' strains.

The distribution of I sequences is different - sibling species show homology to I factor probes and more distantly related species show progressively less, or no, homology (Bucheton et al., 1986). Thus it appears that I sequences are old components of the Drosophila genome which have degenerated in most species. Probes containing certain internal fragments of I factor DNA have been shown to distinguish potentially active I factors from defective I elements (Bucheton et al., 1984; see section 1.4). These probes show the four most closely related species D. melanogaster (I strains), D. simulans, D. sechellia and D. mauritiana all contain potentially active I

factors, other species do not. This gives two possibilities for the appearance of active I factors in D. melanogaster fifty years ago (Bucheton et al., 1986). Firstly, an active I factor may have arisen in one of these four sibling species, possibly reconstituted from defective elements, which then spread between the other species. Secondly, the I factors in D. melanogaster may have become inactive after divergence of the species, and active I factors may later have re-invaded from a sibling species, which still contained active I factors. Engels (unpublished) however has suggested that the simplest explanation for the IR system is inactivation of the I factor in D. melanogaster after prolonged laboratory culture, i.e. loss (of function if not of whole elements). Unlike the PM system there are no species without I-homologous sequences so it is probable that I elements arose in a common ancestor of the I-containing species and were passed by vertical rather than horizontal transmission to D. melanogaster. Inactivation of hereditary I factors after several decades of laboratory culture would be a simpler mechanism as it removes the need for reactivation or horizontal transfer and involves only one event, namely inactivation. As yet there is no evidence to prove either theory.

It is clear from the non-reciprocal nature of dysgenic crosses that two components are involved - a chromosomal component contributed by the male parent and a cytoplasmic component contributed by the female parent. The male factors were found to be linked to any one of the four chromosomes and are now known to be the I and P transposable elements. Inheritance of I and P factors is Mendelian providing these

elements are maintained in I and P strains respectively. When chromosomes carrying these elements are introduced into the interacting cytoplasm (i.e. a dysgenic cross), the elements can move. Chromosomes from the R or M strain may acquire elements and behave as I or P chromosomes in further crosses. This acquisition of I or P potential is known as "chromosomal contamination" (Picard, 1976). Bingham et al. (1982) and Pelisson (1981) have shown that chromosomal contamination is due to the acquistion of I and P factors respectively.

The maternal regulatory component, known as M cytotype in the PM system and reactivity in the IR system, is complex. Firstly, there are strong and weak R strains, and within an unselected strain flies will not be equal in their ability to cause dysgenesis. (This is also seen for I strains, but not to the same extent). With selection at each generation, uniform strong or weak strains may be obtained which will be stable (Picard & L'Heretier, 1971). Some variation is also seen within P and M strains. Secondly, although the strength of interaction depends primarily on the cytoplasm of the R or M strain, ultimately this is dependent upon the genotype. This was demonstrated in the IR system by Bucheton and Picard (1978). One set of lines was constructed which had the chromosomes of a strong R strain and the cytoplasm of a weak R strain. A second set of lines was constructed with the reverse composition. Initially the lines behaved in dysgenic crosses as the strain from which the cytoplasm was derived. However, after several generations (sometimes ten or more were necessary) the strength of interaction switched to that of the strain contributing the

chromosomes. This change is gradual, the level of reactivity varying at each generation. The switch from R to I following a dysgenic cross is not gradual, but occurs in one generation (Picard, 1978). The reactive state cannot be maintained in the presence of chromosomes carrying I factors, whether they are from an I strain or from an R strain which has been contaminated (Picard, 1978).

In the PM system the change of cytotype from M to P may take several generations (Engels, 1979a). At each generation the population as a whole will show a gradual change but each individual fly will be either M or P. Cytotype switch is an all-or-nothing event and at each generation a variable proportion of the flies will switch. The efficiency of cytotype switching is affected by temperature (Ronsseray et al., 1984). A high developmental temperature (26.5°C) of the F_1 flies from the dysgenic cross will promote switching of M to P cytotype in a greater proportion of their offspring than when development is at 18°C. In addition, M to P switching is promoted when females are aged at 28.5°C. A low aging temperature, however, can reverse the effect of a high developmental temperature.

The degree of GD sterility is also sensitive to developmental temperature. Below 25°C hybrids are usually fully fertile. By 29°C sterility is at a maximum (Engels & Preston, 1979; Kidwell & Novy, 1979). High temperature also increases the frequencies of male recombination and transmission ratio distortion.

High temperature and aging affect SF sterility. Old SF females

tend to be more fertile than young SF females - fertility may reach the same level as that of RSF females from the reciprocal cross (Picard et al., 1977). The numbers of lethal mutations and chromosome non-disjunctions also tend to decrease with the age of SF females when mated (Picard et al., 1978; Proust & Prudhommeau, 1982). High developmental temperatures of SF females (29°C) will reduce the hatching percentage (Bucheton, 1979a,b), but this effect can be reversed by high temperature at oogenesis. In contrast, a low temperature (20°C) at oogenesis will greatly reduce hatching percentage of SF females maintained at 29°C throughout earlier developmental stages (Picard et al., 1977; Bucheton, 1978).

There are many unanswered questions concerning hybrid dysgenesis. Firstly, why are dysgenic events confined to the germ line cells? Some progress towards an understanding of this has been made for the PM system (see section 1.3), but not yet for IR. Secondly, how is sterility induced? Mutations and chromosome rearrangements may be explained by insertion or excision of transposable elements. Rubin et al. (1982) and Pelisson (1981) have correlated mutations of the white locus with insertions of P and I element DNA respectively. Engels and Preston (1984) observed a large number of chromosome rearrangements involving 2, 3, 4 or 5 breakpoints. In most cases the breakpoints occurred at or near P element sequences. It is likely that aberrant excision events involving these P element sequences are responsible for these rearrangements. Induction of sterility, however, cannot be explained in such simple terms.

Engels (1983) has suggested that GD sterility may be due to early loss of germ cells. The period of onset of rapid germ cell division coincides with the temperature sensitive period for GD sterility. During this period males have several more times the number of germ cells than females and hence stand more chance of one or more germ cells surviving to form one (or both) gonads. This could explain why males are less prone to GD sterility than females.

Data obtained by Lavige (1986) suggests that SF sterility may be due to an all-or-nothing event which occurs during oogenesis, perhaps a precise function which may or may not occur, or production of a toxic molecule. These suggestions are very speculative however, and further experimentation is needed to gain more insight into the mechanism.

A further question is whether transposition is conservative or replicative. Conservative transposition involves an element excising from one position and inserting at a different location. Replicative transposition does not require excision - an element is copied and the copy inserts elsewhere in the genome leaving the original element in place. It is difficult to determine which mechanism is being employed. Excision of an element and subsequent gain of an element elsewhere is not proof of a conservative mode unless it can be shown to be the same element which has excised and re-inserted. Likewise an increase in the number of elements in the progeny as compared with the parents is not proof of a replicative mode as conservative transposition of an element from one sister chromatid to another following DNA replication

would appear to be replicative if the chromatid which had lost the element was not recovered in the progeny (O'Hare, 1985). Current opinion favours a replicative mode and there is some evidence to support this. Pelisson (unpublished) has shown that R chromosomes can acquire inducer ability from a chromosome carrying a single I factor; the I-bearing chromosome does not lose its inducer ability nor does the I factor change position. Again, however, this could be explained by excision of an I factor from one sister chromatid following DNA replication, with subsequent loss of that chromatid, and insertion of the I factor into one chromatid of a reactive chromosome. Loss of a chromatid or chromosome could be explained if a breakage occurred as a result of excision of an element. It is thought that I factors cannot excise precisely as an IR-induced mutation of the white locus (due to insertion of an I factor) never reverts to wild type in a second dysgenic cross, but gives more extreme white mutations (Pelisson, 1981). These are due to rearrangements of white sequences (Sved, Lynch, unpublished), and may be caused by abortive excision events. Sequencing data obtained so far (Lynch, unpublished) suggests that the I factor remains intact, so may still be active in these cases. A successful excision (perhaps followed by re-insertion elsewhere) may not be detected in this experiment if excision causes chromosome loss.

Engels and Benz (described in Simmons & Karess, 1985) performed an experiment in which an X chromosome containing several P elements was followed through several generations and the number of P elements counted. Insertions into this X chromosome outnumbered excisions by a factor of 1.5-3.0. This is also taken as evidence of a replicative

mode of transposition. It is known that P elements can excise precisely (Rubin et al., 1982; O'Hare & Rubin, 1983), although this observation cannot totally exclude the possibility that aberrant excision events may sometimes occur and cause chromosome loss.

The control of transposition in I and P strains and how this control is overridden when the elements are introduced into R and M strains, is another unanswered question. There are two possible mechanisms - active repression in I and P strains, or active induction in R and M strains. Current opinion favours the former possibility. Two models have been proposed for the PM system. O'Hare and Rubin (1983) suggested that the P factor might code for a transposase and a regulator. If the regulator positively affected its own production whilst negatively affecting the production of transposase, the net effect would be the inhibition of transposition. This would represent the P cytotype. The M cytotype represents the absence of repressor, hence when P factors are introduced they are able to produce transposase and consequently transpose, until the level of regulator is high enough to switch off transposase production. If a threshold level of regulator is required for the inhibition of transposase production, the all-or-nothing nature of the M to P cytotype switch can be explained.

A second model, proposed by Simmons and Bucholz (1985), involves the titration of transposase by extrachromosomal P elements. This model stems from the observation that P factor activity is lower in M strains that contain defective P elements than in M strains that

contain no P elements at all (see section 1.3). Simmons and Bucholz (1985) suggest that extrachromosomal P elements could be formed by the action of transposase and then bind the transposase preventing further transpositions occurring. Strains with more defective P elements would have more potential for binding transposase than strains with few or no P elements which would therefore be stronger M strains. The M cytotype again represents the lack of positive transposition effectors. This model requires the P factor to code for only one product. Extrachromosomal P elements have not yet been identified.

No models have been proposed for the IR system, but models for the PM system could equally well be applied. There is no reason however to suppose that the same mechanism applies in both cases.

1.3 The P Factor

The first direct evidence that a transposable element family is involved in PM hybrid dysgenesis came from studies of dysgenesis-induced mutations of the *white* locus (Rubin *et al.*, 1982; Bingham *et al.*, 1982). Seven independent mutations were examined and all contained insertions into *white*. These insertions fell into two classes. Two mutations were due to *copia* sequences; the remaining five contained sequences related to each other but unrelated to *copia*. These insertions varied in size from 0.5kb to 1.4kb and were proposed to be P elements. When chromosomes containing these insertions were crossed into an M cytotype the mutations were destabilised and the wild type phenotype was restored. Southern blot

analysis of these revertants showed that the insertions had excised, apparently precisely.

Restriction analysis of these elements showed few similarities, but hybridisation studies revealed all five to be related, suggesting a heterogeneous family of P elements. Using these elements as probes to Southern blots of M and P strains Bingham et al. (1982) showed P strains to have a complex pattern of bands characteristic of dispersed repetitive DNA sequences. There are 30-50 copies per haploid genome. Furthermore there is considerable heterogeneity between different P strains, indicative of transposability. Most M strains show little, if any, homology, except for M strains most recently isolated from the wild. These have many P sequences, but they are assumed to be defective. These are called M', or pseudo-M strains.

Using these short elements as probes to genomic DNA, O'Hare and Rubin (1983) isolated a conserved 2.9kb element, proposed to be the P factor. One such 2.9kb element was cloned and sequenced (O'Hare & Rubin, 1983). This element has 31bp inverted repeats at the ends, (Figure 1.1C), and is flanked by an 8bp direct duplication of the target site DNA. Several shorter elements were also sequenced and found to be derived from the 2.9kb element by internal deletions. The inverted repeat ends were found in all these elements, plus an 8bp target site duplication. The P factor itself has four long open reading frames, all on the same strand. Between them they account for most of the sequence.

Spradling and Rubin (1982) demonstrated that this 2.9kb element is fully functional by cloning the element into a plasmid and injecting the construct into M strain embryos. A visual assay for P factor activity was provided by the sn^{w} mutation (Engels, 1979b) carried by the M strain. This mutation is caused by a head-to-head insertion of two defective P elements into the sn locus, and affects bristle morphology. This allele is highly mutable under dysgenic conditions, mutating to either sn^{e} (extreme) or sn^{+} (pseudo-wild type) when one or other of the P elements is mobilised *in trans* by a P factor. The progeny of injected flies displaying a sn^{e} or sn^{+} phenotype were screened for the presence of P sequences in their genome and were found to contain one to five integrated copies. No inserted plasmid DNA was detected, indicating that the P sequences had integrated *via* transposition, not recombination.

The P factor can now be used as an efficient transformation vector for *Drosophila* (Rubin & Spradling, 1982). Only the ends of a P element are required for mobilisation, as evidenced by the ability of P elements with internal deletions to transpose, in the presence of a P factor. This means that internal sequences can be replaced with foreign DNA and the construct can integrate into the genome *via* the flanking P sequences. A functional P factor must be co-injected to provide transposition functions *in trans*.

It is clear that the P factor encodes its own transposition functions. Karess and Rubin (1984) performed experiments to determine which of the four open reading frames were necessary for

transposition. Utilising restriction sites they constructed four mutant P factors, each of which contained a frameshift mutation within one open reading frame. Each mutant P factor also carried the ry^+ gene, to provide a visual assay when injected into ry^- M strain embryos. Each mutant P was able to integrate into the genome when injected singly as a helper P factor was co-injected, but the helper itself lacked part of one inverted repeat end and hence could not integrate. Further transposition events of the mutated P factors (now in the absence of a helper function) were not observed, indicating that all four open reading frames contribute to the transposition function. Flies containing each single mutant P factor were crossed to generate flies carrying all six pairwise combinations of mutant P factors to see if the mutations could complement one another. P factor activity, as assayed by sn^w destabilisation, was not seen, indicating that all four open reading frames contribute to a single transcript, and that splicing must occur to join the coding sequences together. These splice sites have been identified from the sequence and the ORF0-ORF1 and ORF1-ORF2 splice sites verified by ribonuclease protection assays (Laski et al., 1986).

Karess and Rubin (1984) have identified P factor-specific transcripts of 2.5kb and 3kb. These have a common 5' end but the 3' end is different. It is now thought that the 3kb transcript terminates beyond the end of the P factor, whereas the 2.5kb transcript terminates at a polyadenylation site within P (Laski et al., 1986).

The question of why dysgenic events are restricted to the germ

line is some way towards being understood. The possibility of there being a tissue specific promoter has been eliminated (Laski *et al.*, 1986), for two reasons. Firstly, putting the P factor coding region under the control of a promoter which works in somatic tissue (the hsp70 promoter) does not lead to somatic mosaics. Secondly P factor transcripts were detected in somatic tissue indicating that the P factor promoter can work in somatic tissue. Instead, a tissue-specific splice appears to be responsible. When Laski *et al.* (1986) mapped splice juctions in the 2.5kb transcript, no junction between ORF2 and ORF3 was detected, which would result in translation terminating at the end of ORF2, to give a 66kd protein. Yet Karess and Rubin (1984) showed ORF3 was necessary for transposition. When the ORF2-ORF3 splice was made *in vitro* and the resulting P factor injected, somatic mutations were seen. This suggests that it is the ORF2-ORF3 splice which is necessary for germline-specific transposase activity, though the mechanism for this splice is not yet understood. The protein encoded by this spliced RNA is 87kd. Cell lines transformed with the spliced P factor express the 87kd protein, and P element excision is catalysed (Rio *et al.*, 1986). This is not seen when the cell lines are transformed with the unspliced P factor, which produces the 66kd protein. This strongly suggests that the 87kd protein is the transposase. It is possible that the 66kd protein is the regulator molecule proposed by O'Hare and Rubin (1983).

1.4 The I Factor

The I factor transposable element has been identified from a number of mutations of the *white* locus of *D. melanogaster*. These mutations

arose following a number of IR dysgenic crosses involving several different I and R strains (Pelisson, 1981). Eight white mutations have been isolated, called w^{IR1} - w^{IR8}, and these fall into two classes (Bucheton et al., 1984; Sang et al., 1984). w^{IR1} to w^{IR6} are all due to insertions of apparently identical 5.4kb elements. These flies have some residual eye colour, whereas the remaining two strains w^{IR7} and w^{IR8} have white eyes. These two strains both have deletions of part of the white locus. w^{IR1} is thought to be an insertion of an I factor, as the mutation has never been separated from I factor activity by recombination (Pelisson, 1981).

Genomic DNA from I and R strains has been blotted and probed with I factor sequences (Bucheton et al., 1984). All R strains tested contain many sequences homologous to the I probe (in contrast to many M strains) and the banding pattern is very similar in all cases. These elements are presumed to be defective. I strains show a very similar pattern of banding but contain additional bands which vary in position between strains. These extra bands are believed to be I factor sequences, capable of transposition and hence at different positions in different strains. The constancy of the positions of defective elements implies that, unlike defective P elements, defective I elements transpose very rarely.

A HindIII/PstI fragment internal to the I factor has been found to be diagnostic of I strains (Bucheton et al., 1984). This fragment when used as a probe to HindIII/PstI-cut genomic DNA should detect a

band of corresponding size if intact I elements are present. R strains have very few, if any, copies of this band, whereas I strains have 10-15 copies per haploid genome. R strains therefore seem to contain only deletion-deficient I elements.

In situ hybridisation of I factor probes to polytene chromosomes of I and R strains reveals another major difference. In R strains all the hybridisation is to the heterochromatin in the centromeric region of the chromosomes. I strains also show hybridisation to the centromere but have about 15 extra sites of hybridisation on the chromosome arms. Again these sites differ between strains (Bucheton et al., 1984). The I sequences on the chromosome arms are thought to be the I factors, and the sequences in the centromeric region the defective I elements. I elements appear to be confined to the heterochromatin, but the reason for this is not understood.

1.5 Other Drosophila Transposable Elements

(1) The copia-like elements

Many different families of copia-like elements have been identified. Although they closely resemble each other in overall structure, there is very little sequence homology between them with the exception of 297 and 17.6 which appear to be related (Kugimiya et al., 1983).

All copia-like elements have long direct repeat sequences at their ends (Finnegan et al., 1978), called LTRs (Figure 1.1A). Although the length of these repeats is usually constant within a family there is much variation between families, from about 250bp to 500bp. Each family is

represented by about 10-50 copies per haploid genome, the numbers varying in different fly strains. Evidence for the transposability of these elements comes from the diverse locations members of each family are found at when different strains are compared by Southern blotting, or by *in situ* hybridisation.

copia itself was one of the first element families of this kind to be discovered, and was found because it is transcribed to given an abundant polyA$^+$ RNA. Several other *copia*-like elements have been discovered in the same way. Some of these RNAs are full length, and have been shown to start and terminate within the LTRs of the element (Flavell *et al.*, 1981; Scherer *et al.*, 1982).

Sequence data, as well as structural organisation, has shown *copia*-like elements to closely resemble the integrated form of vertebrate retroviruses. Both have LTRs separated by several kilobases of DNA. In retroviruses this central region contains open reading frames - usually three, called *gag*, *pol* and *env*. The products of these genes, which are translated into polyproteins, are necessary for replication of the virus genome (reverse transcriptase), integration into the host cell genome, cleavage of polyproteins, as well as structural proteins for the viral particles.

Two *copia* elements have now been sequenced entirely (Mount & Rubin, 1985; Emori *et al.*, 1985) and one 17.6 (Saigo *et al.*, 1984). These also contain open reading frames, and from the predicted amino acid sequences homology has been found with retroviral *gag*,

and *pol* proteins.

Other features of retroviruses have also been found in *copia*-like elements. Retroviruses have a site for binding of a tRNA primer to the single stranded viral RNA. This is necessary for the first DNA strand to be made, by reverse transcription. *Copia*-like elements also have a primer binding site, at a position analagous to retroviruses, near the 5' end of the RNA (Will *et al.*, 1981; Scherer *et al.*, 1982). At the 3' end of the RNA of retroviruses is a purine-rich sequence, thought to be necessary for synthesis of the second DNA strand. This sequence is also found in *copia*-like elements (Will *et al.*, 1981; Scherer *et al.*, 1982). The organisation of the LTRs is also analogous.

Further evidence for the relationship comes from the discovery of *copia* RNA associated with particles (which morphologically resemble retroviral core particles) in tissue culture cells (Shiba and Saigo, 1983). It seems probable that *copia*-like elements transpose *via* a mechanism of reverse transcription of an RNA intermediate, resembling a retroviral life cycle in all features except the formation of infectious particles. Flavell (1984) has found extrachromosomal, circular, *copia* elements which replicate faster than chromosomal DNA, as measured by uptake of label, and extrachromosomal linear *copia* elements which could incorporate label into both strands even in the presence of an inhibitor of cellular DNA polymerase. Arkhipova *et al.* (1984) have found *mdg1* and *mdg3* in the form of DNA/RNA hybrids. All these results are consistent with *copia*-like elements

transposing by a reverse transcriptase mechanism. In addition, Arkhipova et al. (1986) have detected reverse transcription intermediate forms (minus and plus strong-stop DNA) for mdg1, mdg3 and mdg4 in cultured D. melanogaster cells, which are known to be intermediates of retroviral reverse transcription. This is a further indication of the similarities between retroviruses and copia-like elements.

(2) Fold-back Elements

Fold-back elements are characterised by having long inverted repeat ends, (Figure 1.1B), which are able to anneal together to form stem-loop structures. The first family to be studied, the FB family, was isolated by Potter et al. (1980) by fractionating genomic DNA to enrich for snap-back structures. One element was isolated this way (called FB1) and was used to probe a genomic library to isolate further related sequences.

Although related in sequence, the structure of these elements is very heterogeneous. The lengths of the inverted repeats, both between elements and at the ends of an individual element, can be variable. In addition the length of the DNA between the repeats (the "loop" of the stem-loops structres) can also be variable between elements. Some elements have no separating DNA but are entirely composed of repeat sequences (Potter et al., 1980; Truett et al., 1981).

The inverted repeats are composed of tandem arrays of short, direct repeats (Potter, 1982; Truett et al., 1981). At the outer ends of the element these repeats are 10bp long, these are then

expanded to 20bp long repeats separated by variable lengths of A/T rich DNA. This is followed by 31bp repeats in tandem. These 31bp repeats are not identical, there are five main variants which occur in a regular pattern to give a larger repeat of 155bp. This sort of tandem repeat structure is very similar to that found in satellite DNA; it is possible that FB elements have evolved from such sequences.

FB elements have been shown to mobilise pieces of genomic DNA, sometimes as much as hundreds of kilobases (Ising & Block, 1981, 1984). This may occur when two FB elements inserted into the same chromosome are mobilised, taking with them the genomic sequence between them. Such composite elements are called TE elements. Presumably this occurs when the transposition machinery recognises only one end of each element.

The mechanism of transposition of FB elements is not known. It is unlikely to occur via an RNA intermediate as FB elements have none of the features seen in other elements which utilise reverse transcriptase. Also, it seems highly unlikely that several hundred kb of DNA in a TE element could be transcribed into a single, full length RNA intermediate.

(3) Retroposons

These elements are so named because they have the structure that would be expected if a mRNA were reverse transcribed and the product inserted into the genome (Rogers, 1983). Features of these elements are a lack of direct or inverted terminal repeats, and an A-rich sequence at one

(Figure 1.1D)
end. Retroposons have been found in a variety of species, and fall into three categories. Firstly, there are elements such as processed pseudogenes. These are simply reverse transcripts of a mRNA, and as they contain no coding information for a transposase, or a promoter for an RNA polymerase, there is no reason to suppose that once having inserted these elements are capable of further transposition.

The second class are short (a few 100bp), moderately repetitive sequences, called SINEs (Singer, 1982). Mammalian Alu elements are an example of this class (Houck et al., 1979; Jelinek et al., 1980). SINEs are frequently flanked by direct repeats, suggesting they may be transposable. The number of bases duplicated varies from element to element, unlike most other families of transposable elements (Singer, 1982). RNA homologous to some SINE elements has been found (Weiner, 1980; Haynes & Jelinek, 1981; Pan et al., 1981), and it has been suggested that this may be reverse transcribed to form a transposition intermediate (Singer, 1982). SINEs contain RNA polymerase III promoters, hence if an element did transpose it would take the promoter with it, and presumably be capable of further transposition events. No SINE element has yet been seen to transpose.

The third class is composed of long interspersed sequences, or LINEs (Singer & Skowronski, 1985). These have been detected in mice, rats, dogs, humans and primates, at a copy number of thousands or tens of thousands per genome. These elements are several kilobases long, have an A-rich tail at the 3' end and contain long open reading frames. These elements have the capacity to code for their own

transposition functions, and homology to reverse transcriptase has been found in the putative protein sequences (Loeb et al., 1986; Hattori et al., 1986). A 6.5kb RNA homologous to a human LINE has been found in tissue culture cells (Skowronski & Singer, 1985). Few transposition events have been observed involving LINE elements, but again the ability to do so is inferred from their structure.

In Drosophila melanogaster two families of retroposons have been described to date. These are the F elements and G elements. G elements (Di Nocera & Dawid, 1983; Di Nocera et al., 1986) are found mainly in the non-transcribed spacer of rDNA units and have poly A tracts at the 3' end of one strand.

The first F element to be described was called 101F (Dawid et al., 1981). This element is present in about 50 copies in the genome, at different sites in different strains suggesting transposability. Unlike copia-like elements, FBs and P elements, all members of the same family of F elements do not duplicate the same number of bases at the target site upon insertion (Di Nocera et al., 1983). The 5' end of the element is variable in length, a feature which is common to LINE elements (Singer & Skowronski, 1985).

Although a reverse transcription method of transposition is inferred from the structure of these elements, there is no evidence to prove it. 101F has been partially sequenced, but the sequence available is not sufficient to identify an open reading frame which may code for a reverse transcriptase. No polymerase III promoter has

been found in the sequence.

The elements discussed in this chapter are illustrated in Figure 1.1.

1.6 Aims of this Thesis

The aims of this thesis were to sequence a full length, functional, I factor and to sequence the ends of several other I factors, with the intention of elucidating the structure of this transposable element. Of particular interest were the ends of the element (the presence or absence of terminal repeats), duplication of target site DNA upon integration of an element, target site specificity and correlation of genotype with the phenotypic affect of insertion into the *white* gene. In addition, an analysis of the I factor sequence was undertaken to detect open reading frames and to determine the nature of potential I factor-encoded proteins. It was hoped that this might enable a possible transposition mechanism to be proposed and to explain other features known to be associated with the I factor, such as repression of I factor activity in I strains.

Figure 1.1

Diagrammatic representation of four families of transposable elements.

A. copia-like elements. The repeat structure of the LTRs is indicated (U3, R and U5) and the positions of the primer binding site (PBS) and purine-rich region are shown.

B. Fold-back (FB) elements. Stippled boxes represent inverted repeats, open boxes represent the spacer region.

C. P elements. Shaded boxes represent the short inverted repeats. Open boxes represent the internal P element sequence.

D. Retroposons. The 'A' box represents the A-rich tail. Open boxes represent element sequences.

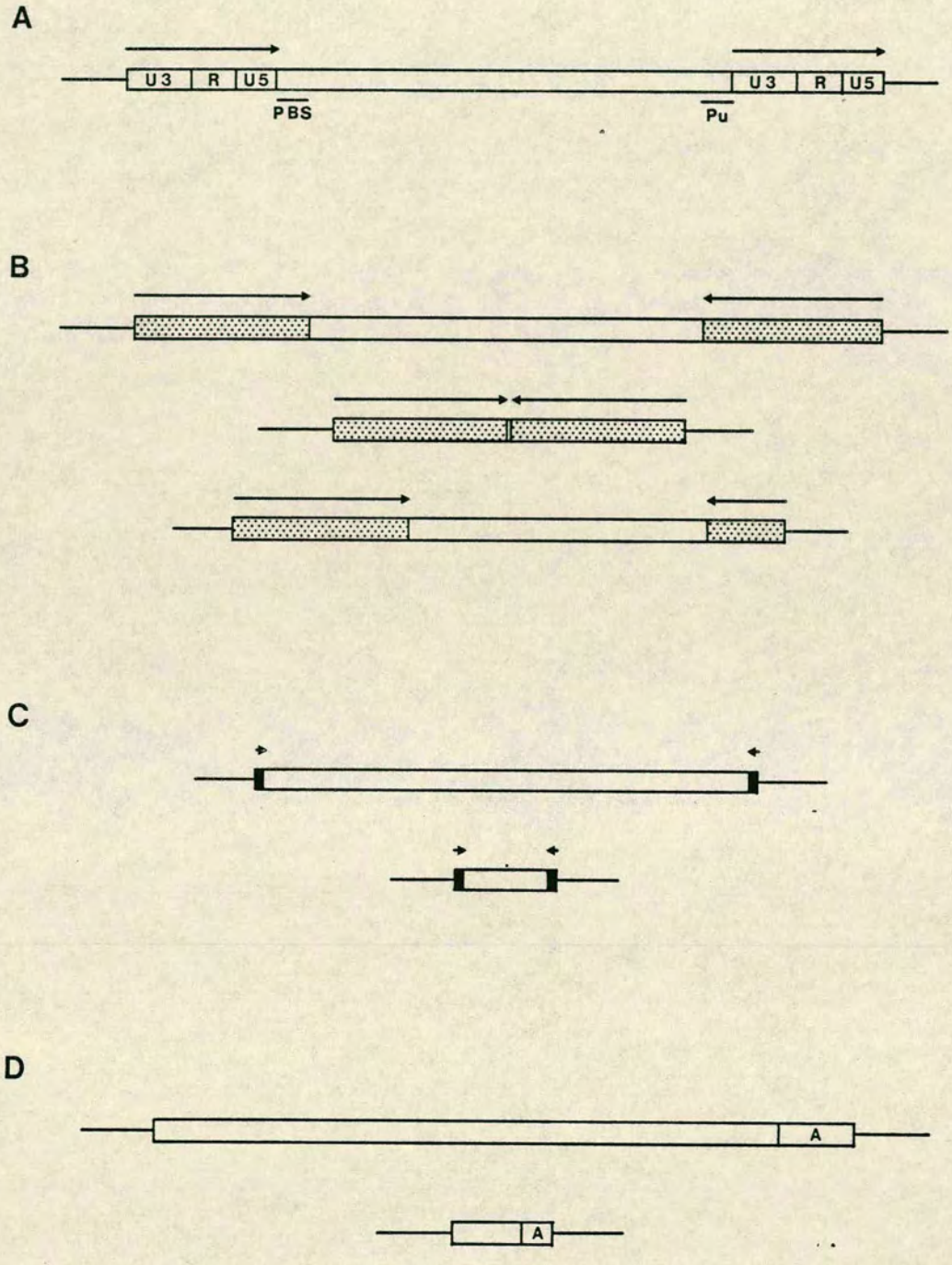


Figure 1.1

CHAPTER 2

Materials and Methods

2.1 Media

L-broth

Difco Bacto tryptone, 10g; Difco Bacto yeast extract, 5g; NaCl, 5g; per litre adjusted to pH 7.2.

L-agar

As L-broth plus Difco agar, 15g/litre.

BBL-agar

Baltimore Biological Laboraties trypticase, 10g; NaCl, 5g; Difco agar, 10g; per litre (pH unadjusted).

BBL top layer

As BBL agar but only 6.5g Difco agar per litre.

Minimal agar

Difco Bacto agar, 6g; 5 x Spizizen salts, 80mls; 20% glucose, 4mls; 5mg/ml vitamin B1, 0.1mls; volume made up to 400mls with water, pH unadjusted.

2.2. Materials

5 x spizizen salts

$(NH_4)_2\ SO_4$, 10g; K_2HPO_4, 70g; KH_2PO_4, 30g; tri-Na Citrate, 5g; $MgSO_4$, 1g; per litre.

Phage buffer

KH_2PO_4, 3g; Na_2HPO_4 (anhydrous), 7g; NaCl, 5g; 0.1M $MgSO_4$, 10mls; 0.01M $CaCl_2$, 10mls; 1% gelatin, 1ml; per litre.

LTB (for storage of M13 plaques)

10mM Tris/HCl pH 8.0; 1mM EDTA

10 x TBE buffer

Tris base, 108g; Boric acid, 55g; EDTA, 9.3g; per litre, pH unadjusted.

20 x SSC

0.3M Na Citrate; 3.0M NaCl.

20 x Denhardts

0.4% bovine serum albumen; 0.4% ficoll (MW 400,000); 0.4% polyvinyl pyrrolidine (MW 40,000).

TFB (for Hanahan competent cells)

10mM K-MES; 100mM KCl; 45mM $MnCl_2.4H_2O$; 10mM $CaCl_2.2H_2O$; 3mM Hexamine cobalt chloride.

Antibiotics

Ampicillin used at 100ug/ml.

Isotopes

α-^{32}P-dCTP (3,000 Ci/mMole) purchased from Amersham International.

^{35}S-dATP (500 Ci/mMole) purchased from New England Nuclear, and from Amersham International (410 Ci/mMole).

Enzymes

Restriction enzymes purchased from Boehringer Mannheim, Pharmacia and Amersham International. E. coli DNA polymerase I purchased from Boehringer Mannheim; Klenow fragment from Boehringer Mannheim and Pharmacia. T4 DNA ligase purchased from New England Biolabs. Calf intestinal phosphatase purchased from Boehringer Mannheim.

4 x NT buffer

210mM Tris/HCl pH 7.5; 21mM $MgCl_2$; 20ug/ml gelatin.

1 x dNTP buffer

100ul 4 x NT buffer; 4ul 2mM dGAT; 1ul 14M β-mercaptoethanol; 295ul H_2O.

TE

10mM Tris/HCl pH 8.0, 1mM EDTA.

TM

100mM Tris/HCl pH 8.0, 50mM $MgCl_2$.

Primers

M13-specific sequencing primers were purchased from New England Biolabs and were of the following sequences:

5' TCCCAGTCACGACGT 3' (15-mer)

5' GTAAAACGACGGCCAGT 3' (17-mer)

M13-specific primer for making single stranded probes was purchased from New England Biolabs:

5' CACAATTCCACACAAC 3'

Insert-specific primers, obtained from Biosearch Incorporated, were of the following sequences:

5' ACGTCTTTGCCACGA 3' (DF1)

5' TTGAATGCTATGAGG 3' (DF2)

5' GCTGTAAGCCCCGTA 3' (DF3)

5' CGCGAAGCGAGGCTT 3' (DF4)

Insert-specific primers obtained from C.J. Leaver, Botany Department, University of Edinburgh, were of the following sequences:

5' CGGTGCTCAAGGAAC 3' (DF5)

5' ATGGTTGCCATCTTG 3' (DF6)

5' GCAGTGGACGCCAGA 3' (DF7)

2.3 Bacteria and Bacteriophages

Phages

Insertion vector λNM1149 (Murray, 1983), for cloning 0-11kb HindIII or EcoRI fragments. Replacement vector λNM762 (Murray et al., 1977;

Williams and Blattner, 1980), for cloning 3-16kb HindIII fragments.

M13mp8 and mp9 (Messing and Vieira, 1982).

M13mp10 (Messing, 1983).

M13mp18 and mp19 (Norrander et al., 1983).

Bacteria

NM514 (hsd R, lyc 7): for growth of recombinant λNM1149 (Murray, 1983).

ED8654 (hsd R, met B, sup E, sup F): for growth of recombinant λNM762 (Borck et al., 1976).

NM430 (hsd R, lac Z amber): for visual selection of recombinant λNM762 (Frischauf et al., 1983).

SM32 (lon$^{\nabla}$, gal E, sul A, str A): for better growth of non-recombinant λNM1149 (Misusawa & Ward, 1982).

NM522 (lac-pro, hsd MS, F'lac ZM15, lacI^{q}); host for M13 phages (Gough & Murray, 1983).

2.4 Drosophila strains

The strains of D. melanogaster described in this thesis arose from the following crosses:

strain	(R) parent	(I) parent	references
w^{IR1}	seF_8	w^1 ct f	Pellison (1981)
			Bucheton et al. (1984)
w^{IR2}	"	"	"
w^{IR3}	"	Luminy	"
w^{IR4}	XCha	XOre I	Sang et al. (1984)
w^{IR5}	"	"	"
w^{IR6}	"	"	"
w^{IR7}	"	"	"
w^{IR8}	"	"	"
bx^{F31}	spontaneous mutation (non-dysgenic)		Peifer and Bender (1986)

2.5 Methods

DNA restriction

All restrictions were carried out in high, medium or low salt buffers as recommended in "A Manual for Genetic Engineering: Advanced Bacterial Genetics", ed. R.W. Davis et al. The volume of enzyme added never exceeded 10% of the total volume. Reactions were stopped by heating at 65°C for 10 minutes or by phenol extraction and ethanol precipitation.

Ligation

Appropriate amounts of vector and donor fragments were incubated overnight at 10°C, with 100 units T4 DNA ligase in a volume of 10ul. Sticky-ended fragments were ligated in 66mM Tris/HCl pH 7.2, 1mM

EDTA, 10mM $MgCl_2$, 10mM DTT, 0.1mM ATP. Blunt-end ligation buffer contained 10-fold higher ATP. Ligations were stored at 4°C.

Agarose gel electrophoresis

DNA was routinely electrophoresed through 0.7-1% Miles agarose in 1 x TBE buffer, at 20 volts (0.77 v/cm) overnight or 100 volts (3.85 v/cm) during the day. Gels were stained after electrophoresis with ethidium bromide and were viewed and photographed using a UV transilluminator.

Recovery of DNA from agarose gels

DNA was electrophoresed through high grade (Seakem) agarose, in the presence of 0.5ug/ml ethidium bromide. When the desired fragment was well separated it was excised and the gel slice placed in dialysis tubing containing 1 x TBE. The DNA was electroeluted at 100 volts for 1 hour, and the current reversed for 1 minute to remove DNA from the wall of the tubing. DNA was recovered either by dialysis of the tubing contents for 1 hour against 1 x TE, followed by ethanol precipitation, or by passage of the TBE through a DE52 column with subsequent phenol extraction and ethanol precipitation of the eluate.

Nick translation (Will et al., 1981).

Approximately 0.5ug DNA was incubated with 10-20uCi α-^{32}P-dCTP, 1ul 2 x 10^{-5} mg/ml DNAse and 1 unit DNA polymerase I, in 20ul 1 x dNTP buffer. The reaction proceeded for 2 hours at 15°C, and was stopped by phenol extraction. Unincorporated nucleotides were separated on a Sephadex G-50 column.

Plaque hybridisation (Benton & Davis, 1977)

Plaques were transferred to nitrocellulose discs by laying the disc on top of the agar. DNA was denatured by soaking the filter in denaturation buffer (0.5M NaOH; 1.5M NaCl). The filters were neutralised in 0.5M Tris, 3.0M NaCl, pH 7.0, washed in 2 x SSC, and baked at 80°C under vacuum. Prehybridisation in 50% formamide, 4 x SSC, 1 x Denhardts was carried out for 30 minutes at 37°C, and the filters were then hybridised overnight in fresh buffer plus heat denatured probe and sonicated calf thymus DNA (25ug/ml) at 37°C. Filters were washed in 2 x SSC, 0.1% SDS for 1 hour at 37°C, and in 2 x SSC for 1 hour at room temperature.

Southern blotting (Smith & Summers, 1980)

DNA was depurinated by soaking the gel in 2 volumes 0.25M HCl for 2 x 15 minutes, and denatured by soaking the gel in 2 volumes 1.5M NaCl, 0.5M NaOH for 2 x 15 minutes. The gel was neutralised in 2 volumes 1M NH_4 acetate, 0.02M NaOH for 2 x 20 minutes. To transfer the DNA, the gel was inverted and a piece of nitrocellulose soaked in 1M NH_4 acetate, 0.02M NaOH placed on top, plus 3 soaked pieces of blotting paper. A stack of dry blotting paper was added and weighted to ensure even contact. Following transfer overnight the filter was washed in 2 x SSC and baked at 80°C under vacuum.

Filters were prehybridised for 2 hours in 2 x SSC, 50% formamide, 1 x Denhardts, 0.1% SDS, 3% dextran sulphate, and hybridised overnight in fresh buffer plus heat denatured probe. Filters were washed twice for 45 minutes in 2 x SSC, 0.1% SDS at 37°C, and twice for 45

minutes in 2 x SSC at room temperature.

M13 single stranded probes (Hu & Messing, 1982)

5ul template was annealed with 2ng primer pHM235, plus 1ul 0.1M DTT, 1ul H_2O and 2ul 5 x HindII buffer (50mM Tris/HCl pH 8.0, 300mM NaCl, 33mM $MgCl_2$, 30mM β-mercaptoethanol, 0.5mg/ml gelatin), at 65°C for 15 minutes. To this was added 5uCi α-^{32}P-dCTP, 1ul 0.5mM dGAT and 1 unit klenow polymerase. Incubation was for 90 minutes, at 15°C, and then the reaction stopped by the addition of 1ul 250mM EDTA, on ice.

1ul of each template to be screened was spotted on a nitrocellulose grid. The filters were prehybridised for several hours at 65°C in 5 x SSC, 5 x Denhardts, 0.1% SDS, 125ug/ml sonicated calf thymus DNA (denatured), and hybridised overnight in the same buffer (but with 1 x Denhardts), plus probe (not denatured) filters were washed for 4 x 1 hour in 0.5 x SSC, 0.1% SDS at 65°C.

Autoradiography

Signals were detected using Dupont Cronex-4 X-ray film, at -70°C with intensifying screens for ^{32}P and at room temperature for ^{35}S. The film was preflashed to an OD_{540} of 0.15 for ^{32}P signals only, (Laskey and Mills, 1977).

Plating cells

For plating λ phage an overnight culture of the host strain was diluted 1:20 in L-broth and grown at 37°C to OD_{650} 0.5. Cells were spun down and resuspended in half the culture volume of 10mM $MgCl_2$.

For plating M13 phage, either an overnight culture of NM522 was used, or a fresh culture grown to at least OD_{650} 0.5. For visual selection of M13 recombinants, 30ul X-gal (20mg/ml) and 20ul MTG (24 mg/ml) were added (per plate) to the plating cells before use. Recombinant plaques are white, non-recombinant blue.

λ plate lysates (Arber et al., 1983)

A single fresh plaque was picked into 0.5ml phage buffer. 0.1ml was mixed with 0.2ml fresh plating cells, and plated on fresh L-plates in top agar. Four plates were usually set up from the same picked plaque. Following incubation at 37°C for 6-7 hours, 4mls L-broth was added to each plate and the plates left overnight at 4°C. To harvest the phage, the L-broth and top layers were removed and pooled, vortexed and centrifuged. The supernatant was removed and stored at 4°C with a few drops of chloroform to inhibit bacterial growth. Titres obtained were routinely 10^{10}-10^{11} phage/ml.

λ liquid lysates (Arber et al., 1983)

A 200ml culture of the host strain in L-broth + 10mM $MgCl_2$ was grown at 37°C to OD_{650} 0.5, and infected with phage at an MOI of 0.1-1. The culture was grown until massive cell lysis occurred, any cells remaining were lysed with chloroform to release phage particles. Cell debris was spun out. The supernatant was treated with DNAse and RNAse to remove bacterial nucleic acids, and the phage precipitated with PEG 6000 overnight at 4°C. Phage were pelleted by centrifugation, resuspended in phage buffer, and banded on a CsCl step gradient, using steps of 1.3, 1.5 and 1.7g/ml CsCl. Phage were removed from the

gradient by side puncture. For preparation of vector DNA the phage were subsequently banded on a 1.5g/ml CsCl equilibrium gradient.

DNA was recovered by phenol extraction to remove phage protein coats, dialysis against TE and ethanol precipitation.

Preparation of plasmid DNA (Will et al., 1981)

A 500ml culture of the plasmid-carrying strain was grown (with ampicillin selection) in L-broth plus 0.2% glucose overnight at 37°C. The cells were pelleted, resuspended in 25% sucrose, 50mM Tris/HCl pH 8.1, 40mM EDTA and 1ml 10mg/ml lysozyme added to digest the cell wall. Lysis occurred upon addition of 13mls triton mix (2ml 10% triton x -100, 25mls 0.5M EDTA, 10mls Tris/HCl pH 8.1, in 100mls). Cell debris was removed by centrifugation and the supernatant phenol extracted and ethanol precipitated. The pellet was resuspended in TE, treated with RNAse and the DNA banded on a 1.55g/ml CsCl density gradient plus ethidium bromide. The lower, plasmid band was removed by side puncture and the ethidium bromide removed by repeated isobutanol extractions. The DNA was recovered by dialysis to TE and ethanol precipitation.

Preparation of M13 RF DNA

One M13 plaque was picked into 1ml L-broth and grown at 37°C for 3 hours. At the same time a single NM522 colony was picked into 10ml L-broth and grown for 3 hours at 37°C. 0.5ml of the plaque culture and 5ml of the NM522 culture were mixed into 500ml L-broth and grown at 37°C overnight. RF DNA was subsequently isolated by the same

procedure as for plasmid DNA.

Preparation of Drosophila DNA

Flies which had been frozen and stored at -80°C were homogenised, on ice, with 5mls 0.025M Tris/HCl pH 7.0, 0.05M EDTA. Cells were lysed by the addition of 5mls phenol, 0.5mls 10% sarkosyl. Cell debris was centrifuged out and the aqueous phase phenol extracted again. Phenol was removed by ether extraction. The DNA was banded on a 1.7g/ml CsCl gradient and removed by side puncture, followed by dialysis to TE and ethanol precipitation.

in vitro packaging (Scherer et al., 1981).

1ug of lambda DNA was mixed with extracts from a prehead donor strain and a packaging protein donor strain, and incubated at 25°C for 1 hour. A further aliquot of packaging protein donor strain extract was added for a further hour. 0.5ml phage buffer was added and the packaged phage stored at 4°C.

Sonication of DNA (Deininger et al., 1983).

30ul of DNA was sonicated in a sonicating water bath for 1 minute bursts with 30 second intervals to prevent heating of the DNA. The DNA was sonicated until all unsonicated DNA disappeared, as assayed previously by running samples from each time point on an agarose gel. For cloning, the ends of the fragments were repaired (see below) and fragments of length 200bp-1000bp were isolated by electroelution from the smear of fragments obtained when the sonicated DNA was run on an agarose gel.

End-repair of DNA fragments

Sonicated fragments with 5' overhanging ends were repaired by incubation of the DNA with 10 units of DNA polymerase I overnight at 10°C, in 10mM Tris, 5mM $MgCl_2$ and 0.025mM dGATC. Sticky ended restriction fragments (with 5' overhanging ends) were incubated with 0.25 units of Klenow polymerase at 37°C for 2 minutes in 20mM Tris/HCl pH 8.0, 7mM $MgCl_2$. dGATC was then added to a concentration of 0.025mM and incubated for a further 10 minutes.

Competent cells

An overnight culture of NM522 was diluted 1:40 in L-broth and grown to OD_{650} 0.35-0.45. The cells were pelleted and resuspended in 1/3 the volume 50mM $CaCl_2$, and placed on ice for 15 minutes. The cells were pelleted again and resuspended in 1/10 the original volume in $CaCl_2$.

For more efficient transformation competent cells were made by the Hanahan modification (Hanahan, 198). Cells were grown to OD_{650} 0.6-0.8, placed on ice for 1 hour, pelleted and resuspended in 1/3 volume TFB. After 15 minutes on ice the cells were spun and resuspended in 1/12.5 volume TFB. DMSO was added (7ul/200ul cells; after 5 minutes DTT was added (2.25M DTT in 40mM KAC, pH 6.0)(7ul/200ul cells) and after 10 minutes a further aliquot of DMSO was added.

Transformation

0.2mls competent cells was mixed with 2-5ng M13, and left on ice for 30 minutes. The cells were heat shocked at 42°C for 90 seconds, to

take up the DNA, and returned to ice. To plate, 0.2mls plating cells plus XGal and MTG was added to each tube and plated in BBL-top layer.

M13 sequencing templates (Sanger et al., 1980)

An overnight culture of NM522 was diluted 1:40 in L-broth and grown to OD_{650} 0.3. M13 plaques were toothpicked into 1ml aliquots of the culture and shaken for $4^{1}/2$ hours at 37°C. The cells were pelleted out and the phage precipitated from the supernatant with PEG 6000. Protein coats were removed with phenol and the DNA ethanol precipitated. Template DNA was resuspended in 40ul 10mM Tris/HCl pH 8.0, 0.1mM EDTA.

Clone-turn around

20ul template was mixed with 2.5ul TM and 2.5ul sequencing primer, and incubated at 60°C for 1 hour to anneal the primer. To this was added 1ul 0.1M Tris/HCl pH 8.0, 1ul 0.1M DTT, 5ul 0.5mM dGATC, 5 units of Klenow and 7ul H_2O. Following incubation at 37°C for 20 minutes, 5ul 0.5mM dGATC and 5 units of Klenow were added and incubated for a further 20 minutes. This incubation was sufficient for the insert to have become double stranded. The DNA was ethanol precipitated and then restricted with two enzymes that cut at either end of the insert. The insert was then cloned into an M13 vector having the same sites but in the opposite orientation.

DNA sequencing (Sanger et al., 1980)

8ul template was mixed with 1ul sequencing primer and 1ul TM, and incubated at 60°C to anneal the primer. The annealed template was

dispensed into 4 x 2ul amounts in capless eppendorf tubes. 2ul of the appropriate termination mix was added to each tube (see below for description of termination mix) plus 2ul Klenow mix (0.4 units Klenow, 1uCi ^{35}S-dATP in 10mM Tris/HCl pH 8.0, 10mM DTT, per tube). The tubes were incubated for 20 minutes at room temperature. 2ul of chase (0.25mM dGATC) was added to each tube, and the tubes incubated a further 20 minutes. For loading on the gel, 2ul of formamide dye was added to each tube (xylene cyanol and bromophenol blue in deionised formamide/10mM EDTA).

Termination mixes

Four termination mixes were used, one for each of the G, A, T and C reactions. Each mix contained the three cold deoxynucleotides and one of the four dideoxy nucleotides. The ratio of the deoxy nucleotide to the equivalent dideoxy nucleotide in each termination mix was calibrated to ensure that termination occurred at every nucleotide within the range of the sequencing gel.

Polyacrylamide gel electrophoresis of sequencing reactions (Biggin et al., 1983)

The DNA samples were boiled for 4 minutes to denature the DNA and then loaded on to a 380mm x 180mm x 0.4mm 6% polyacrylamide gel containing a 2.5 x/0.5 x TBE buffer gradient. The samples were electrophoresed at 25 watts. The gels were fixed in 10% methanol, 10% acetic acid for 15 minutes, transferred to blotting paper and dried down prior to autoradiography.

Compilation of sequence data

Sequence was assembled from individual gel readings using the DBSYSTEM (Staden, 1982) package of computer programs.

Analysis of sequence data

DNA and protein sequences were analysed using the programs written by Devereux et al. (1984) contained in the UWGCG package (version 5). Comparisons of protein sequences and the NBRF data base were performed by J. Collins and A. Lyall using the 'Prelate' system (Collins & Coulson, 1986).

CHAPTER 3

The Sequence of the w^{IR1} I Factor

3.1 Introduction

The mutation $\underline{w}^{IR1}$ was isolated following a dysgenic cross between flies of the inducer strain $\underline{w}^{1}$ ct f and the reactive strain $\underline{seF}_{8}$ (Pellison, 1981). $\underline{w}^{IR1}$ flies have brown eyes which are lighter at 25°C than at 20°C. The white gene controls the deposition of pigment in various tissues of the fly, including the eyes, and is located on the X chromosome. The mutation was thought to have arisen in the X chromosome of the reactive parent as it was found to be linked to the wild-type ct and f loci (Pelisson, 1981). This can also be demonstrated by Southern blotting as the restriction patterns of the two parental white genes differ. This is due to the presence of an F-like transposable element in the control region of the white gene in $\underline{w}^{1}$ ct f. This is the $\underline{w}^{1}$ mutation. The restriction pattern of $\underline{w}^{IR1}$ resembled $\underline{seF}_{8}$ in this region of the gene, hence the insertion arose in the $\underline{seF}_{8}$ chromosome. The mutation was associated with I factor activity and could not be separated from inducer ability by recombination. This placed an I factor within 0.02 map units of the mutation and it was proposed that the mutation was due to the insertion of an I factor (Pelisson, 1981). The segment of the chromosome containing the mutation was capable of contaminating reactive chromosomes, hence this I factor was capable of transposition (Pelisson, 1981).

$\underline{w}^{IR1}$ and $\underline{seF}_{8}$ genomic DNA was restricted and the digestion products comapred by Southern blotting. The filters were hybridised with white gene probes (Bucheton et al., 1984). The only difference

seen was in the 0.86kb SalI fragment of the white gene (Fig. 3.1) which in w^{IR1} was replaced by a 6.2kb fragment. This indicated that a 5.4kb insertion had occurred - the putative I factor. Initial attempts to clone this 6.2kb SalI fragment intact proved unsuccessful (Bucheton, 1981), so instead the fragment was cloned in three pieces making use of two HindIII sites internal to the insertion. The DNA was cloned into lambda and subsequently subcloned into plasmid pAT153 (Twigg & Sherratt, 1980). The two ends were cloned as SalI/HindIII fragments forming plasmid pI770 (left end) and plasmid pI771 (right end) (Bucheton et al., 1984). The central HindIII fragment was subcloned from the lambda clone containing the left end of the element as the clone contained a HindIII fragment generated fortuitously by a partial digestion. The central HindIII fragment was subcloned into pAT153 and called pI786 (Bucheton et al., 1984). These plasmid clones (Fig. 3.2) were used to obtain the entire sequence of the element - the presumptive I factor.

3.2 Cloning of the inserts from pI770, pI771 and pI786 into M13

The first sequence data were obtained by cloning the entire insert from each plasmid into M13 vectors. Two vectors were used, mp8 and mp9 (Messing & Vieira, 1982). These vectors contain the same polylinker but in opposite orientation to enable double digest fragments to be cloned both ways round with respect to the binding site for the M13 sequencing primer.

Bucheton et al. (1984) had mapped the insertion point of the I factor by Southern blotting. It was placed approximately 100bp to the

Figure 3.1

Restriction map of the D. melanogaster white gene. The proposed intron-exon structure of the white gene (O'Hare et al., 1984) is indicated, stippled boxes representing exons. The insertion point of the w^{IR1} I factor within the 0.86 kb SalI fragment is indicated by the symbol 'I'. white restriction fragments used as probes for hybridisation are shown.

Symbols:- H, HindIII; B, BamHI; R, EcoRI; S, SalI.

Figure 3.1

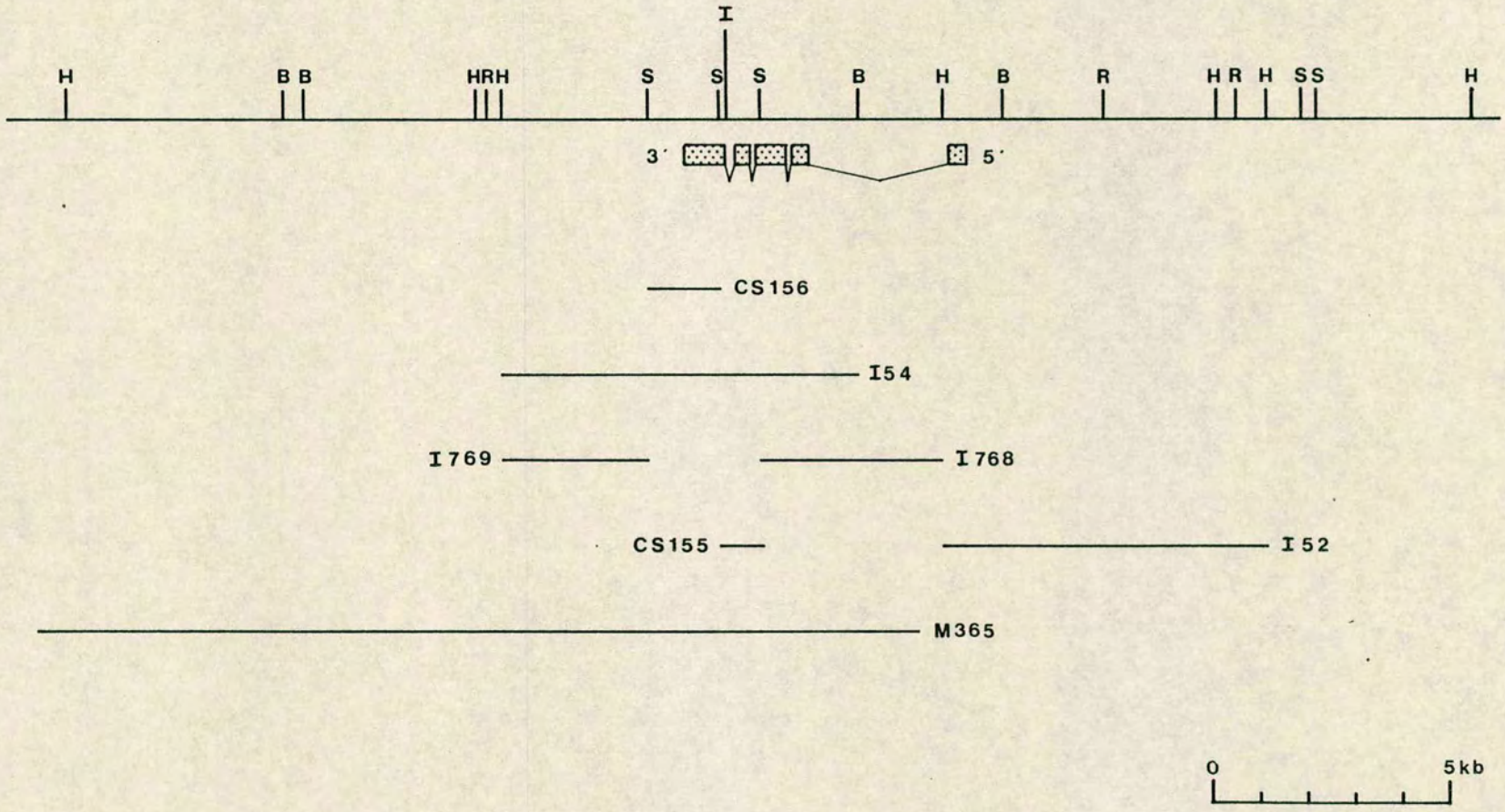
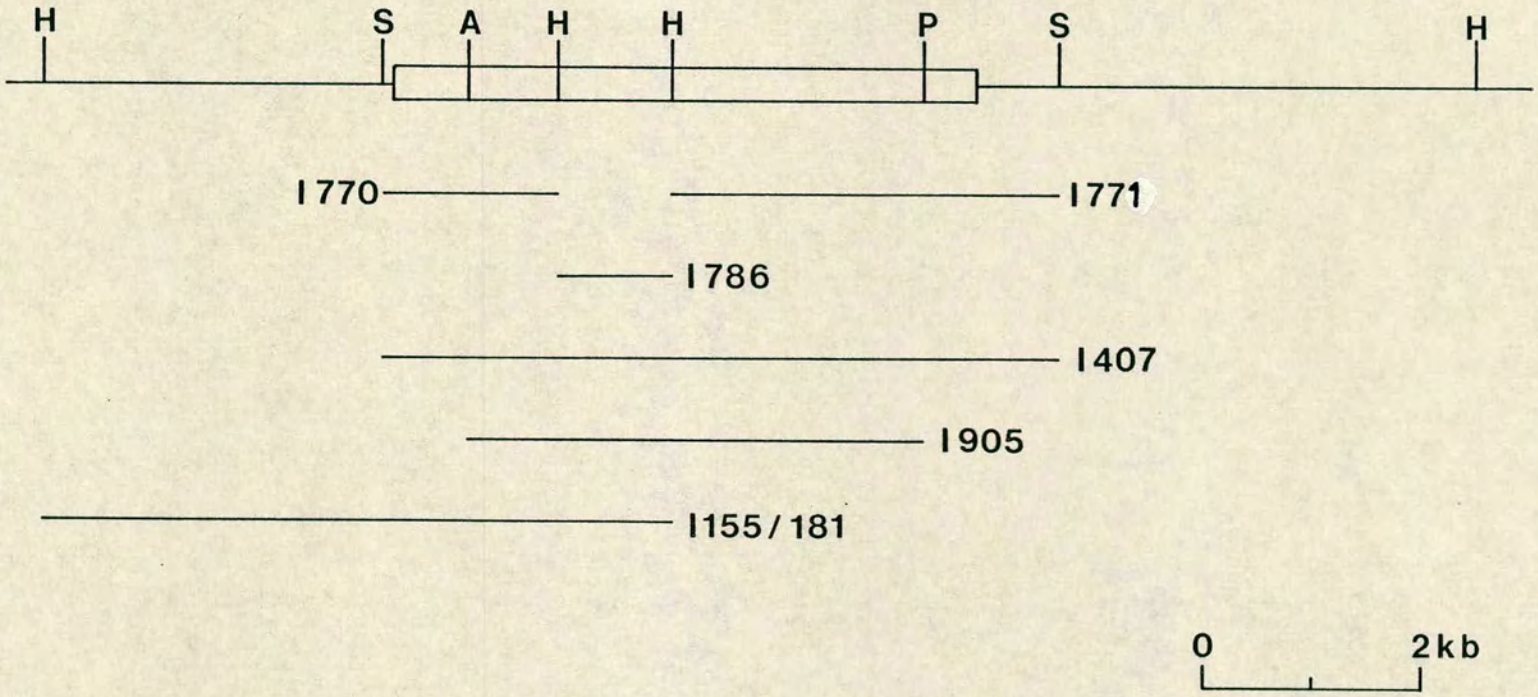

I
H
B B
H R H
S
S
S
B
H
B
R
H R H S S
H
3'
5'
CS156
I54
I769
I768
CS155
I52
M365
0
5kb

Figure 3.2

Phage and plasmid clones containing I factor DNA. All the clones were derived from w^{IR1} with the exception of pI905 and pI407 which were derived from w^{IR3}. The open box represents the I factor inserted in the white gene (single line).

Symbols:- H, HindIII; A, AvaI; S, SalI; P, PstI.

Figure 3.2



right of the left hand SalI site of the 0.86kb SalI fragment (Fig. 3.1). Hence by sequencing from this SalI site of I770 the left end of the I factor could be reached and, by comparison with the wild-type white sequence, the insertion could be mapped exactly. The I770 fragment was therefore cloned in both orientations and sequenced from both ends. The I771 insert was sequenced only from the HindIII site as there was approximately 760bp of white DNA between the SalI site and the right end of the I factor.

Recombinant (white) plaques found after transformation of NM522 were screened with λM365 (Fig. 3.1) to screen out phage carrying pAT153 and several positive phage were picked and sequenced. In addition the 0.86kb SalI fragment of the white gene was cloned from plasmid pCS155 (Fig. 3.1). This fragment is derived from strain Canton S and represents the wild-type white sequence. The fragment was cloned into SalI-cut mp8 and white plaques screened with λM365. Several positives were sequenced and found to contain the insert in both orientations. A comparison with the sequence from the I770 SalI site enabled phage carrying the white fragment in the same relative orientation to be isolated. The insertion point of the I factor was found to be 93bp after the SalI site.

The central 1kb HindIII fragment of the I factor had been purified from the pAT153 vector by preparative agarose gel electrophoresis (R. Paro, this laboratory) and this was ligated into HindIII-cut mp9. Seventeen white plaques were recovered, which were screened with the insert fragment. Only six of the plaques were positive and all these

contained the insert in the same orientation. Although this suggests that the fragment clones preferentially in this orientation six plaques is not a large enough sample to prove this.

3.3 Subcloning of the I factor by sonication

Deininger (1983) described a method for subcloning whereby a large DNA fragment is sonicated to generate shorter fragments (see Chapter 2). This method was used to obtain the bulk of the I factor sequence. It is a random method which relies on a computer to sort the sequences of individual fragments into one contiguous sequence. The programs used for this were part of the DBSYSTEM package, written by Staden (1982).

The cloning of fragments generated by sonication here was found to be an inefficient method. This was probably due to the end-repairing of the sonicated fragments. The DNA will often break to leave 5' or 3' overhangs at one or both ends of the fragment. To generate a clonable fragment these ragged ends had to be removed. DNA polymerase I was used for this. 5' overhangs were repaired by filling in - this was checked by incorporation of a radio-labelled nucleotide into the fragments (data not shown). 3' overhangs should have been removed by the 3' -> 5' endonuclease activity of PolI. When fragments generated by this method were ligated with SmaI-cut mp10 however, religation of the vector tended to be the predominant reaction and white plaques were swamped by blue. Treatment of the cut vector with phosphatase to remove 5' phosphate groups and hence prevent religation overcame this to a large extent, although recovery of white plaques was also reduced. The use of competent cells produced by the Hanahan modification (Hanahan,

1983) enhanced transformation efficiences sufficiently for this not to be a problem. It is possible that the removal of 3' overhangs was at least partly responsible for the inefficient ligation. The method may perhaps be made more efficient by including an S1 nuclease step to remove 3' (and 5') overhanging ends.

White plaques recovered from the cloning of sonicated pI770 and pI786 fragments were screened with λI155 (Fig. 3.1). For pI770 the ratio of negative:positive plaques was expected to be about 2.5:1 as the insert represents just under one-third of the total plasmid. In fact the ratio was 3:1 and this generated sufficient clones to sequence the left end of the I factor well on both strands. To be certain that the sequence was accurate all regions of both strands were sequenced, where possible, at least four times. However, when most of the sequence data had been assembled by the computer it was clear that some regions were represented more frequently than others, leaving short regions where neither strand, or only one strand, was well sequenced. This could be due to the DNA breaking more readily at certain points or to some sequences being less stable than others when cloned.

Templates which contained DNA from a poorly sequenced region were detected by using M13 single stranded probes (Chapter 2). Templates which had been made but not sequenced were spotted on to nitrocellulose and were probed with a template from the well sequenced strand of the same region. Probes were labelled by the method of Hu and Messing (1982).

Figure 3.3

Sequencing strategy for the I factor.

A. Map of the w^{IR1} I factor showing the three plasmid clones (pI770, pI771 and pI786) used for the sequencing.

B. Sequencing strategy. The solid line represents the I factor, flanked by white DNA (single line). Lines above and below the I factor show the extent of plus and minus strand gel readings respectively.

From Fawcett et al. (1986).

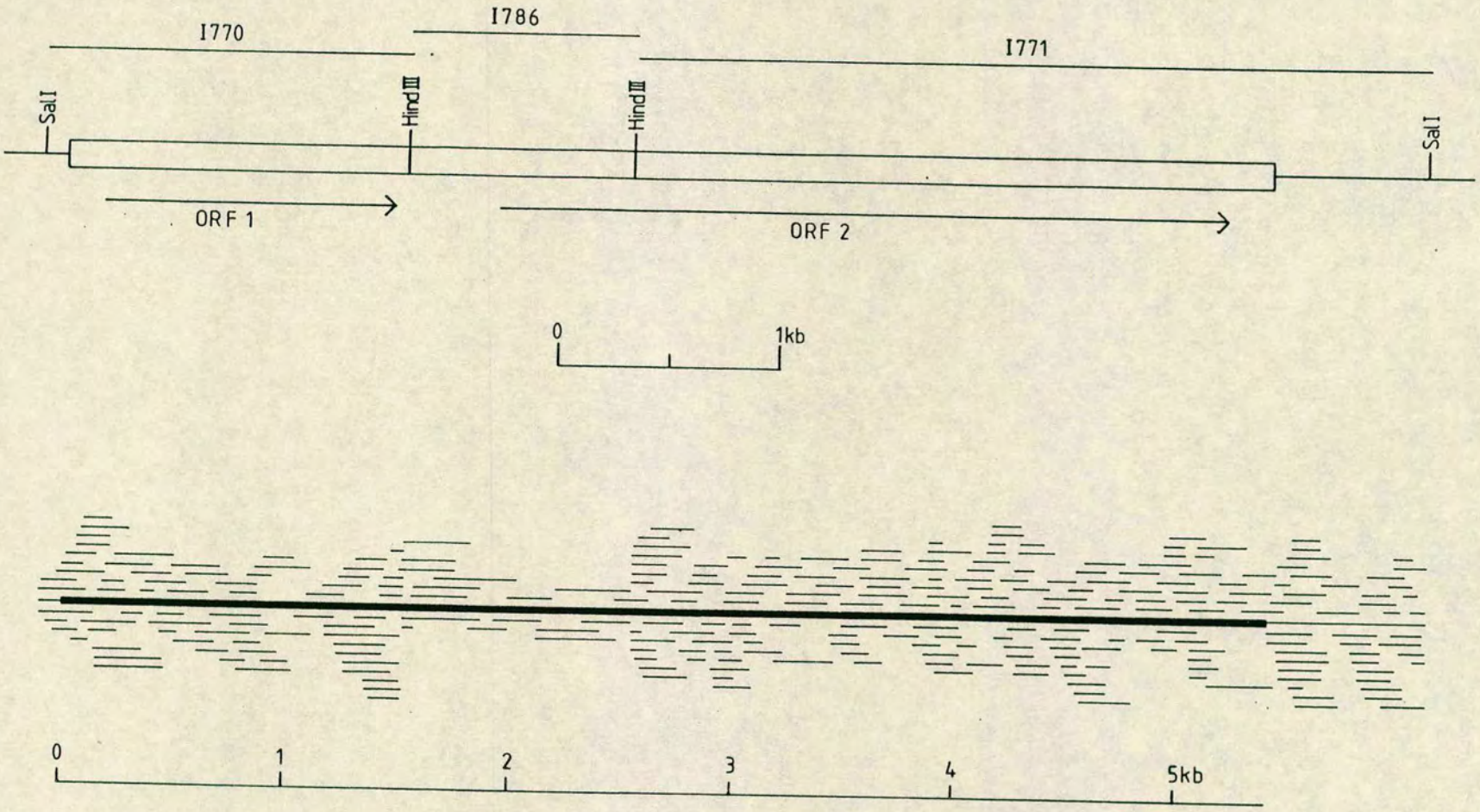

I770
I786
I771
SalI
HindIII
HindIII
SalI
ORF 1
ORF 2
0
1kb
0
1
2
3
4
5kb

Figure 3.3

A few short regions of sequence exist therefore which have been sequenced only once on one strand. This was thought to be adequate providing that the other strand had been sequenced at least once and that the sequences agreed completely.

The sequence of the 1kb HindIII fragment from pI786 was not obtained readily by the sonication method. The recovery of white plaques was always much less efficient than for pI770 and although several attempts were made to overcome this the total number of white plaques recovered was only 85. These were screened with λI155 and although the expected ratio of negative:positive plaques was 3.5:1 only four positives were found. Much of the sequence was determined from these clones (plus the sequence obtained from the intact fragment) but suitable restriction fragments had to be cloned to provide more data and to fill in a gap in the sequence.

The sequence of I771 was determined by C.K. Lister (M.Phil. thesis, 1986). Two regions remained unsequenced on one strand; one region within the I factor sequence and one region spanning the right end of the I factor. These sequences were obtained by myself using oligonucleotide primers (DF1 and DF2, see Chapter 2) synthesised specifically for these regions. The sequencing strategy for the whole I factor is shown in Figure 3.3. The minimum number of times any one base was determined was 2 (i.e. once on each strand). On average, however, each base was determined 6.75 times.

3.4 Sequencing of the I factor internal HindIII fragment

All restriction sites used for the cloning of specific I factor fragments are shown in Figure 3.5.

A gap remained to be sequenced in the I786 DNA. Most of the insert was cut out of the plasmid using restriction enzymes EcoRI and EcoRV (Fig. 3.4). EcoRV cuts once within the I factor DNA (Fig. 3.5). The fragments obtained were ligated into EcoRI/HincII-cut mp18. pI786 was also cut with SalI to prevent the plasmid vector cloning into M13. Only one clonable fragment remained. White plaques obtained were sequenced from the EcoRV site, but the sequence obtained did not bridge the gap.

One of the sonicated clones (called 10/85) looked likely to span the gap as an 8 hour sequencing gel run, from which over 400bp could be read, did not go into M13 sequence. This was the limit of the number of bases which could be read from this clone. To sequence from the other end of the cloned fragment an RF preparation was made of 10/85 (see Chapter 2). The insert was cut out of the RF DNA using EcoRI and SalI which are sites flanking the insert in mp10. This fragment was ligated into EcoRI/SalI-cut mp19, in the opposite orientation. Sequence obtained from these templates started on the opposite side of the gap as expected, but the gap could not be reached even by an 8 hour sequencing gel.

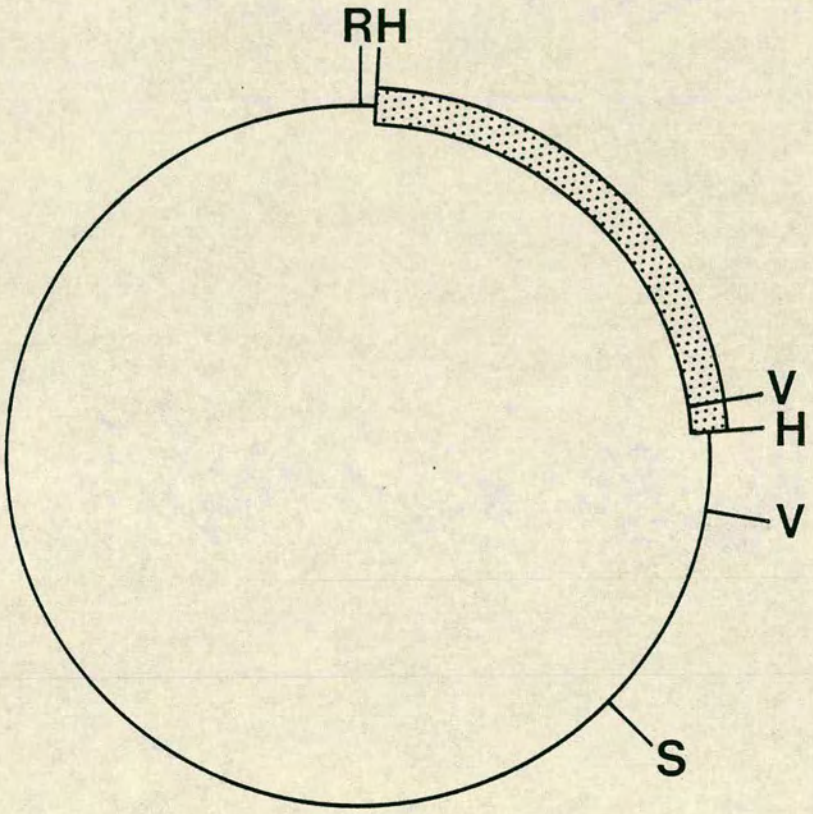
A computer search of the sequence obtained at this stage found one Sau3a site just to one side of the gap (Fig. 3.5). Sau3a/HindIII

Figure 3.4

Plasmid pI786. Stippled box represents I factor sequence; single line represents pAT153.

Symbols:- R, EcoRI; H, HindIII; V, EcoRV; S, SalI.

Figure 3.4



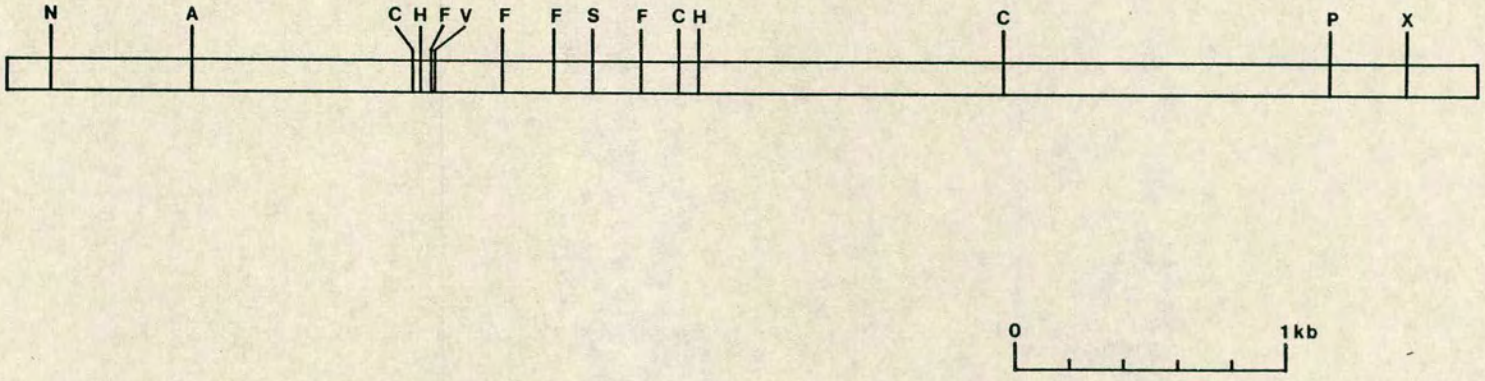
pI786

Figure 3.5

Restriction map of the $\underline{w}^{IR1}$ I factor.

Symbols:- N, HpaII; A, AvaI; C, HincII: F, HinfI; H, HindIII;
V, EcoRV; S, Sau3a; P, PstI; X, XbaI.

Figure 3.5



fragments were cloned from pI786 and probed with λI155. Data obtained from these clones completed the sequence, although additional confirmatory sequence was required. A second computer search, of the complete sequence, found four HinfI sites within this 1kb HindIII fragment (Fig. 3.5). pI786 was digested with HinfI, and the fragments obtained were end repaired (see Chapter 2) and cloned into the SmaI site of mp19. White plaques obtained following transformation were screened with λI155, and positives were sequenced. Data obtained from this experiment completed the sequence of the central HindIII fragment.

3.5 Sequencing over the HindIII sites

Although the complete sequence had been obtained for the three cloned segments of I factor, the orientation of the central HindIII fragment relative to the ends of the I factor was not known. In addition it was necessary to confirm that there are only two HindIII sites in the I factor. Additional sites close to the two used for the initial cloning could generate small fragments which had not been detected. Hence it was essential to obtain sequence data which spanned the two HindIII sites.

As previously described, it had not been possible to clone this I factor intact. However, λI155 contains a HindIII fragment generated by a partial digest of w^{IR1} DNA. This clone contains the left end of the I factor and the central HindIII fragment, plus some white DNA to the left of the I factor insertion point (Fig. 3.2). A computer search of the entire I factor sequence revealed a HincII site in the

I770 sequence 20 bases before the left HindIII site, and another HincII site within the central HindIII fragment. Plasmid pI181 (pAT153 containing the λI155 insert, see Fig. 3.2) was restricted with HincII, and cloned into the HincII site of mp19. White plaques were screened with pI786 and also with pAT153 to eliminate clones containing plasmid sequences. Plaques positive only with pI786 were picked, and templates made and sequenced. From the data produced it was possible to orient the central fragment, and no extra HindIII sites were found in this region.

It was not possible to sequence over the right hand HindIII site of the w^{IR1} I factor, as no clone existed which contained both the right hand end and the central fragment of the I factor. However, another I factor had been cloned, from strain w^{IR3} (see Chapter 1), and this had been cloned intact from the white locus, into pAT153, to form plasmid pI407 (see Fig. 3.2). This I factor, which will be discussed in more detail in Chapter 4, was inserted into the same 0.86kb SalI fragment as the w^{IR1} I factor, and had been cloned as a 6.2kb SalI fragment. pI407 was restricted with HincII and cloned into the HincII site of mp19. White plaques were screened with pI786 and pAT153, and positives with pI786 only were picked and sequenced. The result of this established that there are no extra HindIII sites in this region of the I factor.

EX BIBL. UNIV. EDINBURGEN.

3.6 Features of the w^{IR1} I factor

The I factor is 5371bp long. The sequence of the entire element, plus flanking white DNA, is shown in Figure 3.6. The first feature to be

Figure 3.6

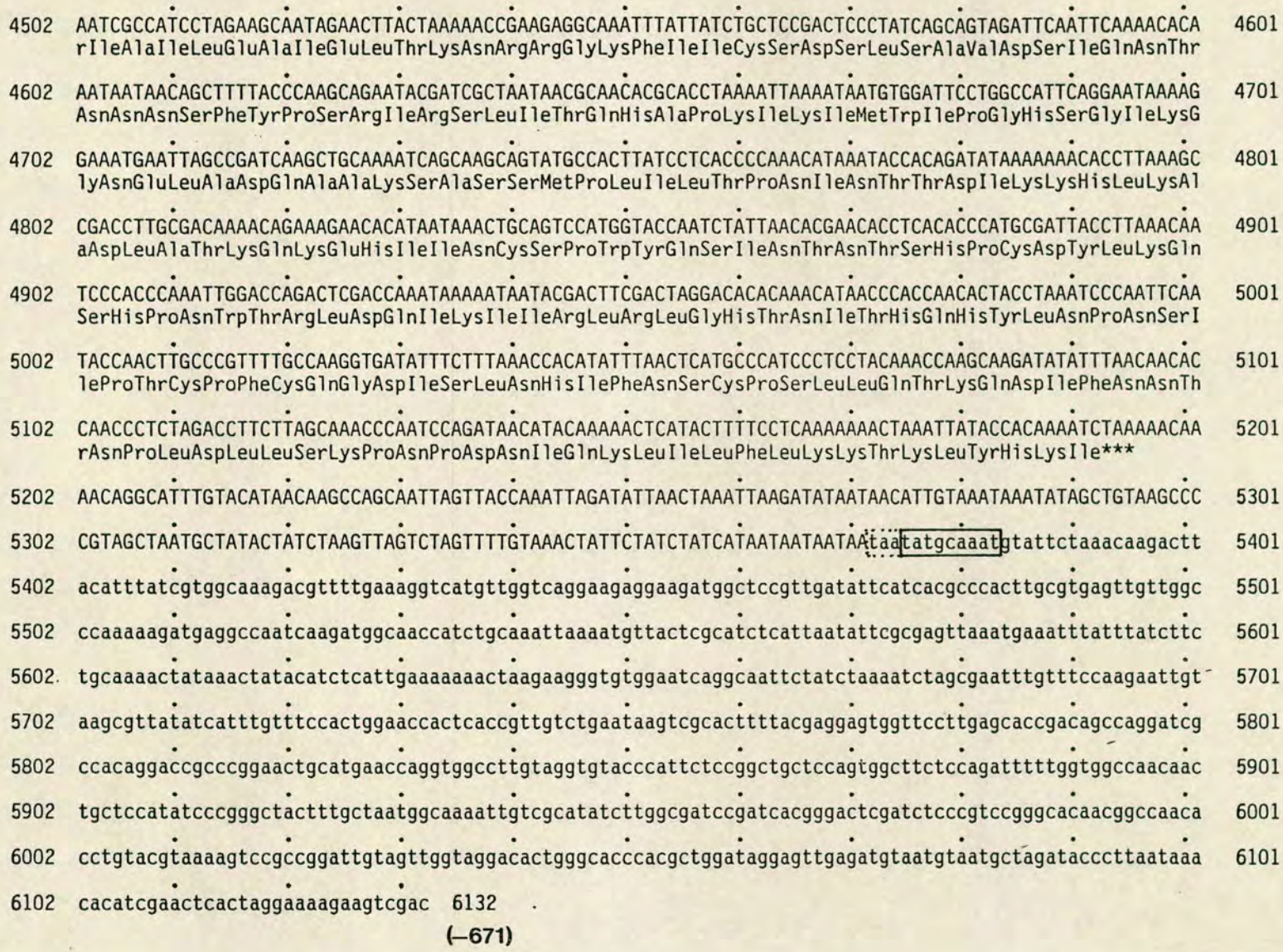
Sequence of the w^{IR1} I factor. I factor sequence is shown in upper case; flanking *white* DNA sequence is shown in lower case. The *white* sequence is the reverse complement of bases -671 to -1535 of O'Hare *et al*. (1984). The target site duplication is boxed. The dotted line shows the TAA triplet which could be either part of the duplication or part of the I factor. The amino acid sequences of the two long open reading frames are shown below the DNA sequence. From Fawcett *et al*. (1986).

(–1535)

```
-99  gtcgacttcgggcctccctcataaaaactggcagctctgaggtgaacacctaaatcgaatcgattcattagaaagttagtaaattattaatatgcaaatc   1

  2  ATTACCACTTCAACCTCCGAAGAGATAAGTCGTGCCTCTCAGTCTAAAGCCTCGCTTCGCGTAAGCCCAAAACTCTTATCAGCAAAATCTTGATAAACAA   101

102  ATATCAACCACAAAGAGAAAATAAAAAACTTAACAACAAAAACAACAATACCGCTAATCCGGGCTCAAGCCCTTAACCAACAATCATGACAGACCCACCA   201
                                                                                    ProThrIleMetThrAspProPro

202  AACATTTACAAAATCACTTCAAAAACATACCAATCCCAATTAGGCGAACCTAAATTTATAATTATTAAAAGAAATGACAACAACTCTTTCGAAAGAACTT   301
     AsnIleTyrLysIleThrSerLysThrTyrGlnSerGlnLeuGlyGluProLysPheIleIleIleLysArgAsnAspAsnAsnSerPheGluArgThrS

302  CACCATTCATCATAAAAAAATCGGTGGACTTTGCCTGTGGAGGAGAAGTTGAGGGATGCAAACGTACAAGAGACGGCAACCTGCTAATAAAAACCAAAAA   401
     erProPheIleIleLysLysSerValAspPheAlaCysGlyGlyGluValGluGlyCysLysArgThrArgAspGlyAsnLeuLeuIleLysThrLysAs

402  TGAATTACAAGCCAGAAAACTCCTAAAACTAACAAAAATTGCAGATGAGGATGTAACAGCAAGTGAACATAAAACATTAAACTTCTCTAAGGGAGTTATT   501
     nGluLeuGlnAlaArgLysLeuLeuLysLeuThrLysIleAlaAspGluAspValThrAlaSerGluHisLysThrLeuAsnPheSerLysGlyValIle

502  TACTGTAACGACCTTAGACACATCGACGAAGACACAATTCTACAAGAACTAAAACCACAAAAAGTATCTGAAGTTAAAAAAATAATGAAACGGCAAAACC   601
     TyrCysAsnAspLeuArgHisIleAspGluAspThrIleLeuGlnGluLeuLysProGlnLysValSerGluValLysLysIleMetLysArgGlnAsnP

602  CCAACTCTAACTCCGACACCAACAACATCACATTAGTTGAAACTGGACTCATAATTATAACCTTTGAATCGCATAAGCTCCCCGAGATAGTACGAATCGG   701
     roAsnSerAsnSerAspThrAsnAsnIleThrLeuValGluThrGlyLeuIleIleIleThrPheGluSerHisLysLeuProGluIleValArgIleGl

702  GTACGAAACAGTCCGAGTACGAGACTATATCCCACTCCCACTTCGATGCAAAAAATGCCTCCGCTTCGGTCATCCAACACCCATATGCAAAAGTGTAGAA   801
     yTyrGluThrValArgValArgAspTyrIleProLeuProLeuArgCysLysLysCysLeuArgPheGlyHisProThrProIleCysLysSerValGlu

802  ACTTGCATCAATTGCTCTGAAACAAAACACACAAACGACGGAGAAAAATGCACAAACGAAAAAAACTGCTTAAATTGCCGAAATAACCCAGAACTTGACC   901
     ThrCysIleAsnCysSerGluThrLysHisThrAsnAspGlyGluLysCysThrAsnGluLysAsnCysLeuAsnCysArgAsnAsnProGluLeuAspH

902  ATCAACACAGCCCAATTGACCGCAAATGCCCTACGTTCATAAAAAACCAGGAATTAACAGCAATTAAAACCACACAAAAAGTTGACCATAAAACGGCCCA   1001
     isGlnHisSerProIleAspArgLysCysProThrPheIleLysAsnGlnGluLeuThrAlaIleLysThrThrGlnLysValAspHisLysThrAlaGl

1002 ACACATATATTTCGAACGTCACGGCTTCCAAACGAAAAACACCTACGCCAAAACACTTACAAACGGCACAACCCAGAGGACAACAAACACTCCATCACCT   1101
     nHisIleTyrPheGluArgHisGlyPheGlnThrLysAsnThrTyrAlaLysThrLeuThrAsnGlyThrThrGlnArgThrThrAsnThrProSerPro

1102 AATATTCACACAAACACAACCCAATCACAACAACAAAATCCGCACCACACACCCAAATCAGCAGCACAAAACACTTCAGCTAAGACACCAACAACTGAAC   1201
     AsnIleHisThrAsnThrThrGlnSerGlnGlnGlnAsnProHisHisThrProLysSerAlaAlaGlnAsnThrSerAlaLysThrProThrThrGluP

1202 CAGCCAAAACAACCTTACTATCCAACCAACCACACCAACACCACCACCACCACAGCTACGACAAACTAGAAGACATGGATACCGACTACACACCTACCAG   1301
     roAlaLysThrThrLeuLeuSerAsnGlnProHisGlnHisHisHisHisHisSerTyrAspLysLeuGluAspMetAspThrAspTyrThrProThrAr

1302 AAAACCATCTACGACATACTCATCACAACTCACAGAAGACCTAAAAATAAAAATCTTCCCTAAAGATAAGTCCAATAACCTATCCATAAACCTTAAAGCA   1401
     gLysProSerThrThrTyrSerSerGlnLeuThrGluAspLeuLysIleLysIlePheProLysAspLysSerAsnAsnLeuSerIleAsnLeuLysAla
```

```
1402 TCAAAACTAAAGGCCAAAGCCCACAAAAACAAGCACACTAACAACAGCGACAGCGAATCCATATAGAACTCTACACAAAACCCTAACCGTTAACACTACC 1501
     SerLysLeuLysAlaLysAlaHisLysAsnLysHisThrAsnAsnSerAspSerGluSerIle***

1502 TTTAAGTAAGTTATAAGCTTTAATTTTCTCACAAATGTCCCTAACTATAATCCAATGGAATCTAAAAGGATATCTAAACAACTACAGCCATCTCCTTATT 1601

1602 CTAATCAAAAAATACTCCCCCCACATAATTTCCCTCCAAGAAACCCATATACAATACACTAATAACATTCCAACCCCAATAAACTACAAACTATTAACAA 1701

1702 ATATTGCCACCAACAGATTTGGGGGGCGTACGACTACTAGTGCATAAGTCAATACAACACACTGTCCTCAACATAACAATCGATATAGAAGCAATAGCCA 1801

1802 TAAATATAGAATCTAAACTTAAATTAAACATATTTTCCACATACATTTCTCCGACCAAAAACATAACTAACCAGACACTCCATAACACATTTAACATACA 1901

1902 ACAAACACCCTCTCTAATTACGGGAGATTTTAATGGATGGCACCATCCTGGGGCTCCCCAACAACAAATAAACGAGGAAAAATAACTCATAGATTCATTG 2001
                                    TrpMetAlaProSerTrpGlySerProThrThrAsnLysArgGlyLysIleThrHisArgPheIleA

2002 ACAACATGCACCTTATCCTGTTAAACGACAAATCTCCCACACACTTTTCAACACACAATACATACACACACATAGACCTCACACTCTGCTCTCCAATCCT 2101
     spAsnMetHisLeuIleLeuLeuAsnAspLysSerProThrHisPheSerThrHisAsnThrTyrThrHisIleAspLeuThrLeuCysSerProIleLe

2102 AGCCCCCCACGCCAAGTGGAAAATACTAAACGATCTTCACGGTAGCGACCATTTCCCTATTATCACAACACTATTCCCAACAACCAATCCACAAAAATTC 2201
     uAlaProHisAlaLysTrpLysIleLeuAsnAspLeuHisGlySerAspHisPheProIleIleThrThrLeuPheProThrThrAsnProGlnLysPhe

2202 TACAGACCCTTTTTTAAACTCAAAGAAGCCAACTGGGAACAGTTCAACGCTCTTACCCACCAAACCAACAAGAAATACCCCACCTCCCACAACGTAAACA 2301
     TyrArgProPhePheLysLeuLysGluAlaAsnTrpGluGlnPheAsnAlaLeuThrHisGlnThrAsnLysLysTyrProThrSerHisAsnValAsnL

2302 AAGAAGCCGCTCTAATCAATAGAATCATCCTTTATAGCGCAAACCTCTCCATCCCACAAACCTCACCTAACACACATCCATACAGGGTTCCATGGTGGAA 2401
     ysGluAlaAlaLeuIleAsnArgIleIleLeuTyrSerAlaAsnLeuSerIleProGlnThrSerProAsnThrHisProTyrArgValProTrpTrpAs

2402 TAAACACCTCGACCAATTACGTAAAGAAAAACAACTTGCCTGGAAAAAATTAAACCGCACAATTACTGTTGACAACATTCTAGACTATAGACGCAAAAAC 2501
     nLysHisLeuAspGlnLeuArgLysGluLysGlnLeuAlaTrpLysLysLeuAsnArgThrIleThrValAspAsnIleLeuAspTyrArgArgLysAsn

2502 GCAATATTTAGATACGAACTAAAAAAGAGGAAAAAAGAAGCTTCCAGCTCTTTCACCTCAACCATCCATCCCACTACTCCCTCATCCAAAATATGGGCCA 2601
     AlaIlePheArgTyrGluLeuLysLysArgLysLysGluAlaSerSerSerPheThrSerThrIleHisProThrThrProSerSerLysIleTrpAlaA

2602 ATATAAGACGCTTCTGCGGACTTAACCCAGCAAAACAAATTCATGCCATCACAAACCCAGTAAATAACGAGACTACATTGGCTAGCAACGAAATTGCTAA 2701
     snIleArgArgPheCysGlyLeuAsnProAlaLysGlnIleHisAlaIleThrAsnProValAsnAsnGluThrThrLeuAlaSerAsnGluIleAlaAs

2702 CATATTCGCACAACATTTCTCTGACCTCTCCGGCGACTGGAACTTCTCAGAGGAGTTCCGGAACAATAAATATAGAAATAACATACATCTCTACACCCCC 2801
     nIlePheAlaGlnHisPheSerAspLeuSerGlyAspTrpAsnPheSerGluGluPheArgAsnAsnLysTyrArgAsnAsnIleHisLeuTyrThrPro

2802 TCTCCAATAGCCCAAACCATAGAAGAGAACATAACGTATCTAGAACTTAGCTCAGCACTACAAACATTAAAAGGATGTGCTCCAGGACTAAATAGAATCT 2901
     SerProIleAlaGlnThrIleGluGluAsnIleThrTyrLeuGluLeuSerSerAlaLeuGlnThrLeuLysGlyCysAlaProGlyLeuAsnArgIleS

2902 CGTATCAAATGATCAAAAATAGCTCCCACACAACAAAAAACCGAATAACGAAACTATTTAATGAAATATTCAATAGCCACATACCTCAAGCCTACAAAAC 3001
     erTyrGlnMetIleLysAsnSerSerHisThrThrLysAsnArgIleThrLysLeuPheAsnGluIlePheAsnSerHisIleProGlnAlaTyrLysTh
```

```
3002 AAGCCTAATCATCCCAATCCTTAAGCCAAACACCGACAAAACGAAAACTTCCTCATACCGACCCATCTCCCTCAACTGCTGTATAGCAAAGATACTTGAT 3101
     rSerLeuIleIleProIleLeuLysProAsnThrAspLysThrLysThrSerSerTyrArgProIleSerLeuAsnCysCysIleAlaLysIleLeuAsp

3102 AAAATAATTGCGAAAAGACTCTGGTGGCTAGTGACATATAACAACCTAATTAACGACAAACAATTCGGGTTCAAAAAAGGCAAATCGACTTCGGACTGTC 3201
     LysIleIleAlaLysArgLeuTrpTrpLeuValThrTyrAsnAsnLeuIleAsnAspLysGlnPheGlyPheLysLysGlyLysSerThrSerAspCysL

3202 TACTCTATGTAGACTATCTCATAACGAAGTCAAAAATGCACACCTCCCTCGTCACTCTTGATTTTTCAAGAGCCTTCGATCGAGTAGGTGTGCACTCCAT 3301
     euLeuTyrValAspTyrLeuIleThrLysSerLysMetHisThrSerLeuValThrLeuAspPheSerArgAlaPheAspArgValGlyValHisSerIl

3302 AATCCAGCAATTGCAGGAATGGAAAACGGGTCCCAAAATAATAAAATACATTAAAAACTTCATGAGCAACAGAAAAATAACTGTCCGCGTCGGTCCGCAT 3401
     eIleGlnGlnLeuGlnGluTrpLysThrGlyProLysIleIleLysTyrIleLysAsnPheMetSerAsnArgLysIleThrValArgValGlyProHis

3402 ACATCAAGCCCGTTACCCCTATTCAACGGAATCCCCCAAGGTTCACCCATATCCGTAATACTTTTCCTCATAGCATTCAACAAATTATCCAACATCATAT 3501
     ThrSerSerProLeuProLeuPheAsnGlyIleProGlnGlySerProIleSerValIleLeuPheLeuIleAlaPheAsnLysLeuSerAsnIleIleS

3502 CCCTACATAAAGAAATTAAATTCAACGCATATGCCGACGACTTCTTCCTTATAATAAATTTCAACAAAAACACAAATACAAATTTCAACTTAGACAATCT 3601
     erLeuHisLysGluIleLysPheAsnAlaTyrAlaAspAspPhePheLeuIleIleAsnPheAsnLysAsnThrAsnThrAsnPheAsnLeuAspAsnLe

3602 ATTCGACGATATAGAAAATTGGTGCTCCTACTCAGGGGCATCGCTTTCCCTATCCAAATGTCAACACCTCCACATATGCAGAAAACGTCACTGCACATGC 3701
     uPheAspAspIleGluAsnTrpCysSerTyrSerGlyAlaSerLeuSerLeuSerLysCysGlnHisLeuHisIleCysArgLysArgHisCysThrCys

3702 AAGATAAGCTGCAACAACTTCCAAATTCCTAGCGTTACGTCCTTAAAAATTCTAGGAATAACCTTAAACAACAAATACAAATGGAACACACACATAAACC 3801
     LysIleSerCysAsnAsnPheGlnIleProSerValThrSerLeuLysIleLeuGlyIleThrLeuAsnAsnLysTyrLysTrpAsnThrHisIleAsnL

3802 TACTTCTACCCAAACTACACAACAAGCTAAATATAATAAAATGCCTATCTAGTCTTAAATTTAACTGCAACACGCATACACTACTTAATGTCGCAAAAGC 3901
     euLeuLeuProLysLeuHisAsnLysLeuAsnIleIleLysCysLeuSerSerLeuLysPheAsnCysAsnThrHisThrLeuLeuAsnValAlaLysAl

3902 AACAATTATAGCCAAACTAGAGTATGGTTTGTTTCTGTACGGCCATGCTCCCAAAAGCATTTTAAACAAAATAAAAACACCGTTTAACTCCGCTATCCGT 4001
     aThrIleIleAlaLysLeuGluTyrGlyLeuPheLeuTyrGlyHisAlaProLysSerIleLeuAsnLysIleLysThrProPheAsnSerAlaIleArg

4002 CTAGCTCTCGGCGCATATCGCTCTACCCCAATAAATAACTTACTTTACGAATCGAATACTCCCCCCTTAGAAATGAAACGAGACCTTCAAATAGCCAAAC 4101
     LeuAlaLeuGlyAlaTyrArgSerThrProIleAsnAsnLeuLeuTyrGluSerAsnThrProProLeuGluMetLysArgAspLeuGlnIleAlaLysL

4102 TATCCCAAAACCTAATCCTCTCCAAAAACACACCAATACATAAGTTCTTAAAGCCTAAAAAAGCTAATAAGAAAAAAACATCAACAATAGACCGAACAAT 4201
     euSerGlnAsnLeuIleLeuSerLysAsnThrProIleHisLysPheLeuLysProLysLysAlaAsnLysLysLysThrSerThrIleAspArgThrIl

4202 CAAACTTAGCCTAGAACTTAATCTACCCTACAAACCAATAAAACTCCATAAAAACAAACCACCATGGACCCTCCCCAATCTAATAGACACGTCACTTAGA 4301
     eLysLeuSerLeuGluLeuAsnLeuProTyrLysProIleLysLeuHisLysAsnLysProProTrpThrLeuProAsnLeuIleAspThrSerLeuArg

4302 ATCCATAAGAAAGAACAAACATCTCCAGACCAATACAGAAAATTATACGAACACACAAAGAATAACCTCAAAACACACAATTTCATATTCACTGACGGTT 4401
     IleHisLysLysGluGlnThrSerProAspGlnTyrArgLysLeuTyrGluHisThrLysAsnAsnLeuLysThrHisAsnPheIlePheThrAspGlyS

4402 CAAAAATTAATTACACAATATCATTCGCCATTACAACGGAGACAGACGTCTTGAAATACGGCATACTGCCCCCATATTCATCCGTCCTCACCTCCGAAAC 4501
     erLysIleAsnTyrThrIleSerPheAlaIleThrThrGluThrAspValLeuLysTyrGlyIleLeuProProTyrSerSerValLeuThrSerGluTh
```


4502 AATCGCCATCCTAGAAGCAATAGAACTTACTAAAAACCGAAGAGGCAAATTTATTATCTGCTCCGACTCCCTATCAGCAGTAGATTCAATTCAAAACACA 4601
rIleAlaIleLeuGluAlaIleGluLeuThrLysAsnArgArgGlyLysPheIleIleCysSerAspSerLeuSerAlaValAspSerIleGlnAsnThr
4602 AATAATAACAGCTTTTACCCAAGCAGAATACGATCGCTAATAACGCAACACGCACCTAAAATTAAAATAATGTGGATTCCTGGCCATTCAGGAATAAAAG 4701
AsnAsnAsnSerPheTyrProSerArgIleArgSerLeuIleThrGlnHisAlaProLysIleLysIleMetTrpIleProGlyHisSerGlyIleLysG
4702 GAAATGAATTAGCCGATCAAGCTGCAAAATCAGCAAGCAGTATGCCACTTATCCTCACCCCAAACATAAATACCACAGATATAAAAAAACACCTTAAAGC 4801
lyAsnGluLeuAlaAspGlnAlaAlaLysSerAlaSerSerMetProLeuIleLeuThrProAsnIleAsnThrThrAspIleLysLysHisLeuLysAl
4802 CGACCTTGCGACAAAACAGAAAGAACACATAATAAACTGCAGTCCATGGTACCAATCTATTAACACGAACACCTCACACCCATGCGATTACCTTAAACAA 4901
aAspLeuAlaThrLysGlnLysGluHisIleIleAsnCysSerProTrpTyrGlnSerIleAsnThrAsnThrSerHisProCysAspTyrLeuLysGln
4902 TCCCACCCAAATTGGACCAGACTCGACCAAATAAAAATAATACGACTTCGACTAGGACACACAAACATAACCCACCAACACTACCTAAATCCCAATTCAA 5001
SerHisProAsnTrpThrArgLeuAspGlnIleLysIleIleArgLeuArgLeuGlyHisThrAsnIleThrHisGlnHisTyrLeuAsnProAsnSerI
5002 TACCAACTTGCCCGTTTTGCCAAGGTGATATTTCTTTAAACCACATATTTAACTCATGCCCATCCCTCCTACAAACCAAGCAAGATATATTTAACAACAC 5101
leProThrCysProPheCysGlnGlyAspIleSerLeuAsnHisIlePheAsnSerCysProSerLeuLeuGlnThrLysGlnAspIlePheAsnAsnTh
5102 CAACCCTCTAGACCTTCTTAGCAAACCCAATCCAGATAACATACAAAAACTCATACTTTTCCTCAAAAAAACTAAATTATACCACAAAATCTAAAAACAA 5201
rAsnProLeuAspLeuLeuSerLysProAsnProAspAsnIleGlnLysLeuIleLeuPheLeuLysLysThrLysLeuTyrHisLysIle***
5202 AACAGGCATTTGTACATAACAAGCCAGCAATTAGTTACCAAATTAGATATTAACTAAATTAAGATATAATAACATTGTAAATAAATATAGCTGTAAGCCC 5301
5302 CGTAGCTAATGCTATACTATCTAAGTTAGTCTAGTTTTGTAAACTATTCTATCTATCATAATAATAATAAtaatatgcaaatgtattctaaacaagactt 5401
5402 acatttatcgtggcaaagacgttttgaaaggtcatgttggtcaggaagaggaagatggctccgttgatattcatcacgcccacttgcgtgagttgttggc 5501
5502 ccaaaaagatgaggccaatcaagatggcaaccatctgcaaattaaaatgttactcgcatctcattaatattcgcgagttaaatgaaatttatttatcttc 5601
5602 tgcaaaactataaactatacatctcattgaaaaaaactaagaagggtgtggaatcaggcaattctatctaaaatctagcgaatttgtttccaagaattgt 5701
5702 aagcgttatatcatttgtttccactggaaccactcaccgttgtctgaataagtcgcacttttacgaggagtggttccttgagcaccgacagccaggatcg 5801
5802 ccacaggaccgcccggaactgcatgaaccaggtggccttgtaggtgtacccattctccggctgctccagtggcttctccagattttggtggccaacaac 5901
5902 tgctccatatcccgggctactttgctaatggcaaaattgtcgcatatcttggcgatccgatcacgggactcgatctcccgtccgggcacaacggccaaca 6001
6002 cctgtacgtaaaagtccgccggattgtagttggtaggacactgggcacccacgctggataggagttgagatgtaatgtaatgctagataccccttaataaa 6101
6102 cacatcgaactcactaggaaaagaagtcgac 6132

(–671)

noted was the lack of direct or inverted terminal repeats, features characteristic of the majority of other transposable element families. Instead, at the 3' end of the top strand (the designated "right hand" end of the element) there is a short stretch of 4 or 5 TAA triplets.

Another feature characteristic of transposable elements is a direct duplication of target site DNA flanking the inserted element. This was identified for the I factor by comparing the white sequences at either end of the I factor and also by comparison with the "empty site" sequenced from the pCS155 clone. This revealed a direct duplication of a 9 or 12bp sequence found only once at the empty site. The reason for the uncertainty over the number of bases duplicated and the number of TAA triplets is because there is a TAA triplet in the white sequence close to the insertion point of the I factor, which may be included in the duplication, giving a 12bp duplication and four TAA triplets. Alternatively the 5th TAA may be part of the I factor, giving a 9bp duplication.

To assess the coding capacity of this I factor, the positions of all methionine codons and stop codons in the three open reading frames of both top and bottom strands were plotted using the UWGCG program FRAMES. This shows the position of all open reading frames (Fig. 3.7). Only two long open reading frames were identified from this, designated ORF1 and ORF2, both on the top strand but in different reading frames. Between them they occupy the bulk of the I factor sequence - ORF1 starts at bases 178 and ends at base 1464 and ORF2 starts at base 1935, ending at base 5192. Both these open reading

Figure 3.7

Output of the UWGCG program FRAMES. Positions of methionine codons (above the lines) and stop codons (below the lines) are shown for the three forward and three backward reading frames of the I factor. ORFs 1 and 2 are indicated.

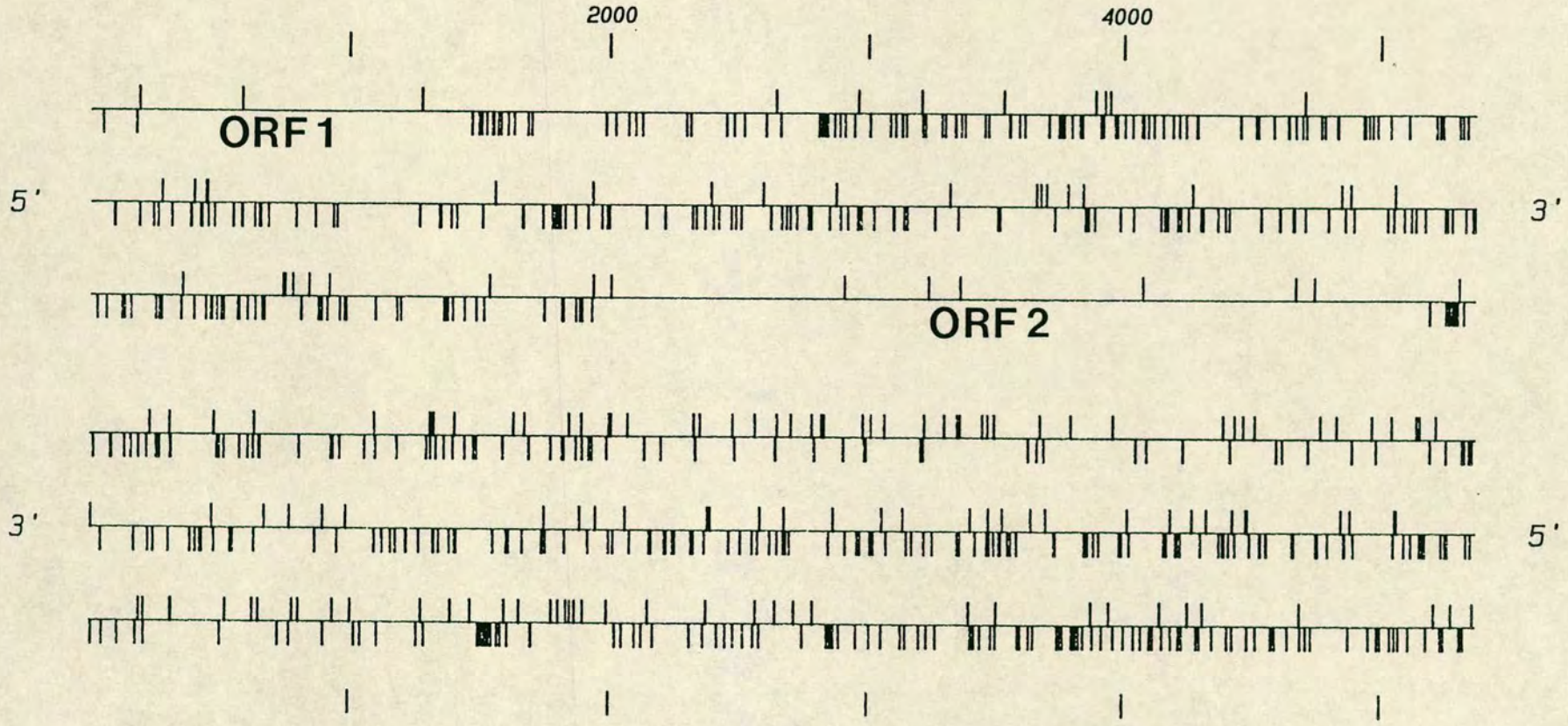

2000
4000
ORF 1
5'
3'
ORF 2
3'
5'
FRAMES of: CON38.SEG Ck: 4918, 1 to: 5371 30-SEP-1986 16:35

Figure 3.7

frames were translated, using the UWGCG program TRANSLATE, to give the peptide sequence of potential gene products (Fig. 3.6).

3.7 Comparison of the I factor with other transposable element families

The lack of terminal repeats would appear to distinguish the I factor from most other transposable element families, such as the copia-like elements (long, direct terminal repeats), fold-back elements (long, inverted terminal repeats) or P elements (short, inverted terminal repeats). However, a closer comparison with the P factor was carried out because of the involvement of both element families in hybrid dysgenesis. This was performed in two ways, firstly by a DNA sequence comparison, and secondly by a comparison between the potential protein products of the I factor open reading frames and those of the P factor (see Chapter 1, section 1.3). Both comparisons employed the UWGCG program WORDSEARCH. By neither method could any homology be found. This is not unexpected however as it is known from genetic studies that the IR and PM systems of hybrid dysgenesis are quite distinct (Kidwell, 1979) and hence there is no reason to expect the two transposable elements to be related.

The remaining class of transposable elements with which to compare I are the retroposons (Rogers, 1985). As previously mentioned (Chapter 1, section 1.5) these elements have no terminal repeats, being characterised instead by an A-rich sequence at the 3' of one strand. The I factor has an A-rich sequence in the form of the TAA triplets, but it does not seem likely that this could have arisen from a poly A sequence by mutation as the pattern is too regular. At this stage it

was unclear as to the relationship, if any, between I and the retroposons, although the I factor clearly resembles retroposons more closely than any other group. This relationship will be discussed in more detail later.

3.8 Analysis of the sequence

In order for an RNA to be efficiently translated, the methionine codon (AUG) must be in a favourable "context". Kozak (1986) has demonstrated that a methionine codon within the sequence ACCATGG has the context which gives optimum translation efficiency. By creating mutations of this sequence by site-directed mutagenesis, the A at position -3 and G at position +4 were found to be most important (Kozak, 1986). The methionine codons near the beginning of ORF1 and ORF2 were examined in the light of these experiments. In ORF1, the first methionine occurs at codon 4, and is contained within the sequence ATCATGA. Although not the optimum context this would still allow efficient translation as there is an A at position -3. Translation from this methionine would result in a protein of 426 amino acids, or 48,000 daltons. In ORF2 there are two methionine codons near the start, at codons 2 and 25. These are contained within the sequences TGGATGG and AACATGC respectively. Neither of these is an ideal context but it is possible that methionine 25 would be used in preference as a T at position -3 in combination with a G at +4 was not found to be particularly efficient (Kozak, 1986). Translation from methionine 25 would result in a 1062 amino acid protein (121,000 daltons); translation from methionine 2 would give a 1086 amino acid protein (124,000 daltons).

Another, essential, requirement for expression of these open reading frames are promoter sequences. The canonical RNA polymerase II promoter sequences ("CAAT" box and "TATA" box, Efstratiadis et al., 1980) were searched for in the sequence preceding ORF1 and ORF2. Before ORF1 potential "CAAT" boxes are found at positions 41 and 69 and potential "TATA" boxes at positions 102 and 122. For ORF2 potential "CAAT" boxes are at positions 1751, 1778 and 1793 and potential "TATA" boxes at positions 1807, 1831 and 1842. The similarities between these sequences and the canonical sequences are rather weak (Fig. 3.8). However, several genes have been identified whose promoter sequences bear a similarly weak degree of homology (see Efstratiadis et al., 1980, for review), so the possibility that these sequences could be used in the I factor cannot be ruled out. The start points of transcription products could be mapped by primer-extension of I factor RNAs. This would require the isolation, from Drosophila, of I factor-specific RNA known to come from an active I factor. This approach may not be possible, however, as such transcripts are proving difficult to detect (Prosser, Lister and Finnegan, unpublished data).

Alternatively, the use of potential promoter sequences could be tested by constructing derivatives of the I factor wherein these sequences had been altered by site-directed mutagenesis. The effects of such mutations could be assayed in an in vitro transcription system or possibly by transformation of Drosophila R strain embyros. The development of an I factor transformation system is underway (Dura, Pritchard, Bucheton and Finnegan, unpublished data).

Figure 3.8

Comparison of potential I factor promoter sequences with known 'CAAT' and 'ATA' boxes from other eukaryotic genes (from Efstratiadis *et al.*, 1980).

CONSENSUS:	CCAAT	ATA
Human Globin	CCAAT	CATAAAA
Human Globin	CCAAT	CATAAAA
Adenovirus Early 1A	TCAAA	TATTTAT
Human Insulin	CCAGG	TATAAAG
Ovalbumin	TCAAA	TATATAT
Drosophila Histone H2A	TCAAT	TATAAAT
Drosophila Histone H3	TCAAA	TATAAGT
I Factor ORF1	TCAGT(41)	AATATCA(102)
I Factor ORF1	CCAAA(69)	AATAAAA(122)
I Factor ORF2	TCAAT(1751)	AATAGAA(1807)
I Factor ORF2	ACAAT(1778)	CATATTT(1831)
I Factor ORF2	GCAAT(1793)	CATACAT(1842)

Figure 3.8

Potential polyadenylation signals (AATAAA) following both open reading frames have been identified. These occur at position 1679 (ORF1) and position 5281 (ORF2).

To detect any DNA sequences with homology to the I factor the DNA sequence was compared against the GENBANK nucleic acid sequence library (release 32.0), using the program WORDSEARCH. The sequences detected as being most similar to I were almost exclusively organelle sequences, mainly mitochondrial and some chloroplast. The degree of homology however was not very high, and the homology tended to be scattered - no more than 10 or 12 bases at the most could be aligned in one region. Organelle sequences tend to be fairly A/T rich. The I factor DNA is 62% AT, which is approximately the same as total Drosophila DNA. The regions where the sequences were aligned tended to be where short runs of A/T residues occur in I factor DNA. It is probable that there is no significant homology between the I factor and organelle sequences and as in the GENBANK database there are no sequences which bear any genuine homology to I, organelle sequences may simply be the least dis-similar. This could be because the best alignment could be achieved between similar kinds of sequence, i.e. A/T-rich.

Often a more sensitive method for comparing sequences is to use protein rather than DNA sequences. This is because in general protein sequences are more highly conserved than DNA sequences. Due to the redundancy of the genetic code two fairly different DNA sequences may

code for similar proteins. In addition, protein comparison programs may allow matches between not only identical but also chemically similar amino acids. This means that proteins which have quite different sequences but similar structures or properties may be detected. The translations of ORF1 and ORF2 were compared individually against the NBRF protein sequence database. As with the DNA sequence comparisons, no protein sequences with any extensive homology to either reading frame were detected. However, if two proteins have a similar function for which only a few amino acids in the protein are functionally required, the overall level of homology between the proteins may be low and the few conserved amino acids would not be detected.

As has previously been mentioned, of all the families of transposable elements the I factor most closely resembles the retroposons in structure. Retroposons, as their name implies, are thought to transpose *via* reverse transcription of an RNA intermediate. (For reviews see Rogers, 1985; Weiner *et al.*, 1986). If the I factor were also a retroposon, one product it may code for is a reverse transcriptase. Toh *et al* (1983, 1985) compared the amino acid sequences of known and putative reverse transcriptases from a variety of retroviruses, the *Drosophila copia*-like transposable element 17.6, cauliflower mosaic virus (CaMV) and hepatitis B virus (HBV). CaMV and HBV, although not retroviruses, are thought to employ a reverse transcription step in their replication cycle. This comparison showed 12 absolutely conserved and 15 chemically similar amino acids in all these proteins, mostly scattered between seven

conserved "domains" within each protein (Michel & Lang, 1985; Hattori et al., 1986). Regions similar to these seven conserved regions have also been found within the open reading frames of some mitochondrial class II introns in fungi (Michel & Lang, 1985). Mitochondrial introns are grouped as class I or class II introns according to their sequence and possible secondary structure. Toh et al. (1983) have suggested that these conserved regions represent the functional regions of reverse transcriptases. The amino acid sequences of ORF1 and ORF2 were searched to see if these same domains were present in the I factor. ORF1 showed no homology but within ORF2 regions corresponding to all 7 domains were found (see Fig. 3.9A). Ten of the 12 conserved amino acids and eleven of the fifteen chemically similar amino acids can be identified. This strongly suggests that the I factor ORF2 does code for a reverse transcriptase, and hence that I is a true retroposon, i.e. transposes via reverse transcription of an RNA intermediate.

The class of retroposons which I most closely resembles are the mammalian LINE elements (Singer, 1982). These elements have been found in several species including mouse, rat, dog, a variety of primates and humans. Each species has one major LINE family, called L1 elements, which are repeated about 10^4 times in the genome (Singer & Skowronski, 1985). Full length LINE elements are 6-7kb long, and have an A-rich sequence at the designated 3' end. The majority of elements are truncated, mostly missing sequences from the 5' end. In addition a few elements have internal deletions and rearrangements. Some, but not all, LINE elements are flanked by a target site duplication, suggesting that these elements may be transposable (Singer &

Figure 3.9A

Comparison of ORF2 of the I factor with known and putative reverse transcriptases. The seven conserved domains (Michel and Lang, 1985; Hattori *et al.*, 1986) are shown. Viral sequences are rous sarcoma virus (RSV), murine mammary tumour virus (MMTV), hepatitis B virus (HBV) and cauliflower mosaic virus (CaMV). The *copia*-like element 17.6 is shown. L1 element sequences are mouse (L1Md-A2), rat (L1Rat), human consensus (L1Hs) and Nycticelus coucang (L1Nc). S.c.-a1 and S.c.-a2 are sequences encoded by introns a1 and a2 of the mitochondrial cytochrome oxidase subunit 1 gene of *Saccharomyces cerevisiae*. Triangles indicate positions at which Toh *et al.* (1985) found identical or chemically similar amino acid residues between eight reverse transcriptases. Filled triangles indicate that the I factor encodes an identical or similar amino acid, an open triangle indicates an unrelated amino acid in the I factor sequence. The filled circle indicates an additional identical amino acid in the I factor and the viral sequences. Boxes show where the I factor and all the LINE elements have identical residues. The number of amino acids separating each domain are shown. Amino acid abbreviations follow the standard single letter code. The following groups of residues with similar properties were used for comparing sequences: P, A, G, S and T (neutral or weakly hydrophobic); Q, N, E and D (hydrophilic, acid amine); H, K and R (hydrophilic, basic); L, I, V and M (hydrophobic); F, Y and W (hydrophobic, aromatic); C (cross-link forming).

From Fawcett *et al.* (1986).

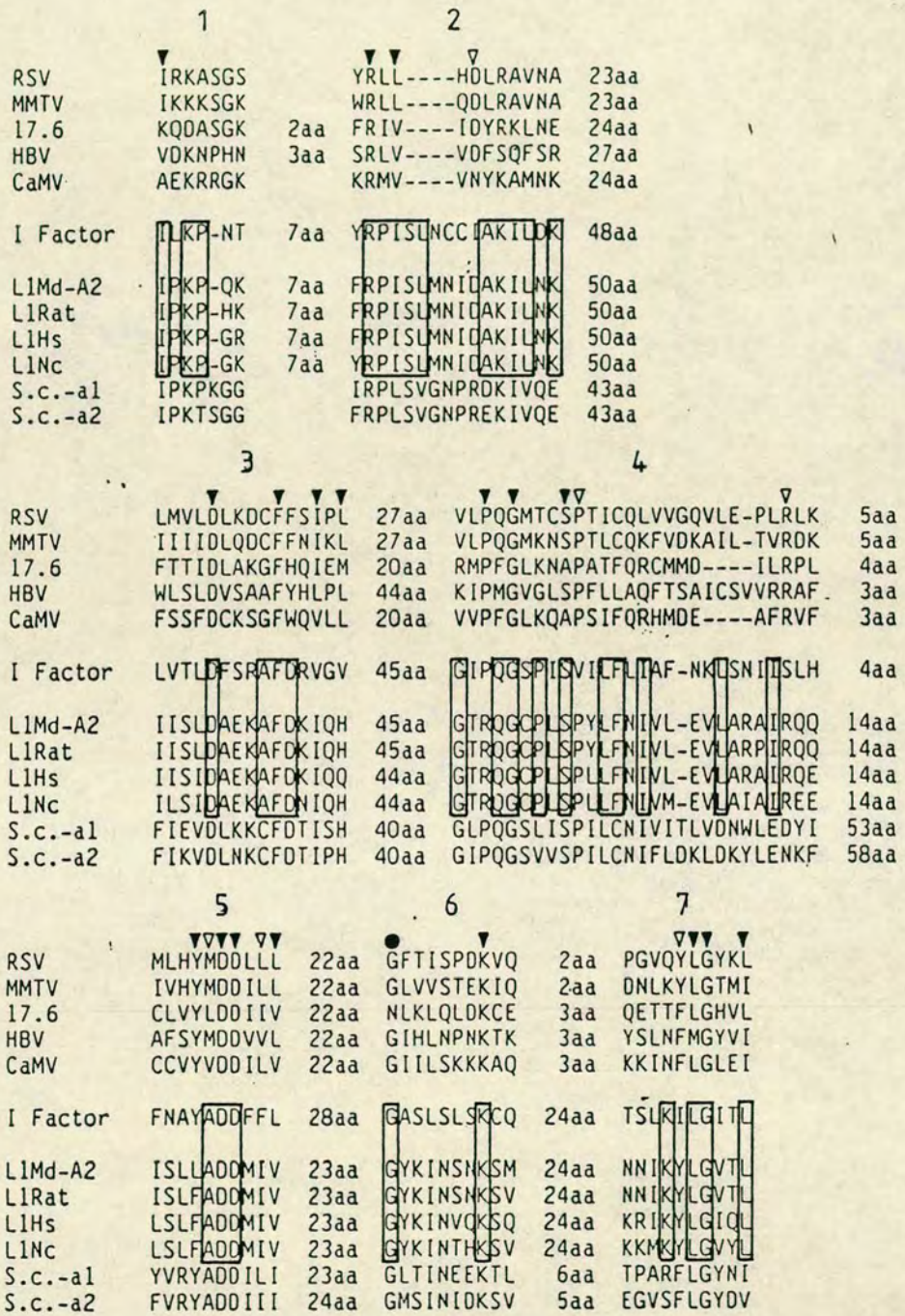

1
2
RSV IRKASGS YRLL----HDLRAVNA 23aa
MMTV IKKKSGK WRLL----QDLRAVNA 23aa
17.6 KQDASGK 2aa FRIV----IDYRKLNE 24aa
HBV VDKNPHN 3aa SRLV----VDFSQFSR 27aa
CaMV AEKRRGK KRMV----VNYKAMNK 24aa
I Factor IIKP-NT 7aa YRPISLNCCIAKILDK 48aa
L1Md-A2 IPKP-QK 7aa FRPISLMNIIAKILNK 50aa
L1Rat IPKP-HK 7aa FRPISLMNIIAKILNK 50aa
L1Hs IPKP-GR 7aa FRPISLMNIIAKILNK 50aa
L1Nc IPKP-GK 7aa YRPISLMNIIAKILNK 50aa
S.c.-a1 IPKPKGG IRPLSVGNPRDKIVQE 43aa
S.c.-a2 IPKTSGG FRPLSVGNPREKIVQE 43aa
3
4
RSV LMVLDLKDCFFSIPL 27aa VLPQGMTCSPTICQLVVGQVLE-PLRLK 5aa
MMTV IIIIDLQDCFFNIKL 27aa VLPQGMKNSPTLCQKFVDKAIL-TVRDK 5aa
17.6 FTTIDLAKGFHQIEM 20aa RMPFGLKNAPATFQRCMMD----ILRPL 4aa
HBV WLSLDVSAAFYHLPL 44aa KIPMGVGLSPFLLAQFTSAICSVVRRAF 3aa
CaMV FSSFDCKSGFWQVLL 20aa VVPFGLKQAPSIFQRHMDE----AFRVF 3aa
I Factor LVTLDFSRAFDRVGV 45aa GIPQGSPISVILFLIAF-NKDSNIISLH 4aa
L1Md-A2 IISLDAEKAFDKIQH 45aa GTRQGCPLSPYLFNIVL-EVLARAIRQQ 14aa
L1Rat IISLDAEKAFDKIQH 45aa GTRQGCPLSPYLFNIVL-EVLARPIRQQ 14aa
L1Hs IISIDAEKAFDKIQQ 44aa GTRQGCPLSPLLFNIVL-EVLARAIRQE 14aa
L1Nc ILSIDAEKAFDNIQH 44aa GTRQGCPLSPLLFNIVM-EVLAIAIREE 14aa
S.c.-a1 FIEVDLKKCFDTISH 40aa GLPQGSLISPILCNIVITLVDNWLEDYI 53aa
S.c.-a2 FIKVDLNKCFDTIPH 40aa GIPQGSVVSPILCNIFLDKLDKYLENKF 58aa
5
6
7
RSV MLHYMDDLLL 22aa GFTISPDKVQ 2aa PGVQYLGYKL
MMTV IVHYMDDILL 22aa GLVVSTEKIQ 2aa DNLKYLGTMI
17.6 CLVYLDDIIV 22aa NLKLQLDKCE 3aa QETTFLGHVL
HBV AFSYMDDVVL 22aa GIHLNPNKTK 3aa YSLNFMGYVI
CaMV CCVYVDDILV 22aa GIILSKKKAQ 3aa KKINFLGLEI
I Factor FNAYADDFFL 28aa GASLSLSKCQ 24aa TSLKILGITL
L1Md-A2 ISLLADDMIV 23aa GYKINSNKSM 24aa NNIKYLGVTL
L1Rat ISLFADDMIV 23aa GYKINSNKSV 24aa NNIKYLGVTL
L1Hs LSLFADDMIV 23aa GYKINVQKSQ 24aa KRIKYLGIQL
L1Nc LSLFADDMIV 23aa GYKINTHKSV 24aa KKMKYLGVYL
S.c.-a1 YVRYADDILI 23aa GLTINEEKTL 6aa TPARFLGYNI
S.c.-a2 FVRYADDIII 24aa GMSINIDKSV 5aa EGVSFLGYDV

Figure 3.9A

Skowronski, 1985). The number of bases duplicated may vary from one element to another.

Loeb et al. (1986) have determined the complete base sequence of a long (6.85kb) LINE element from the mouse, called L1Md-A2. This element contains two long open reading frames of 1137bp and 3900bp, which overlap by 14bp. The protein potentially encoded by the larger open reading frame contains regions homologous to the reverse transcriptase domains (Fig. 3.9A). In addition, the sequences of a rat LINE (L1Rn, d'Ambrosio et al., 1986) and consensus sequences of human/primate LINEs (Singer & Skowronski, 1985) and human/prosimian LINEs (Hattori et al., 1986) have been determined. Regions showing homology to reverse transcriptase have been found within these sequences too. Because both the 3900bp open reading frame of L1MD-A2 and ORF2 of the I factor contain homology to reverse transcriptase, these two protein sequences were compared directly. The strongest homology is found within the region of domains 1 and 2, where 19/34 amino acids are identical and 23/34 are chemically similar or identical. Part of this region is shown in Figure 3.9B where the seven domains of several reverse transcriptases have been aligned. These include retroviruses RSV (rous sarcoma virus) and MMTV (mouse mammary tumour virus), 17.6, fungal class II introns S.c. - a1 and S.c. - a2 (introns a1 and a2 of the mitochondrial cytochrome oxidase subunit 1 gene of Saccharomyces cerevisiae), LINE elements L1Md-A2, L1Rn, L1Hs (human), L1Nc (slow loris) and the I factor. The homology between the various LINEs and the I factor is clearly greater in region 2 than the homology between I (or LINEs) and the other reverse transcriptases,

Figure 3.9B

The highly conserved region between the I factor and L1Md-A2. The positions of domains I and II of reverse transcriptases are shown. Filled circles indicate identical amino acids, open circles indicate chemically similar amino acids.

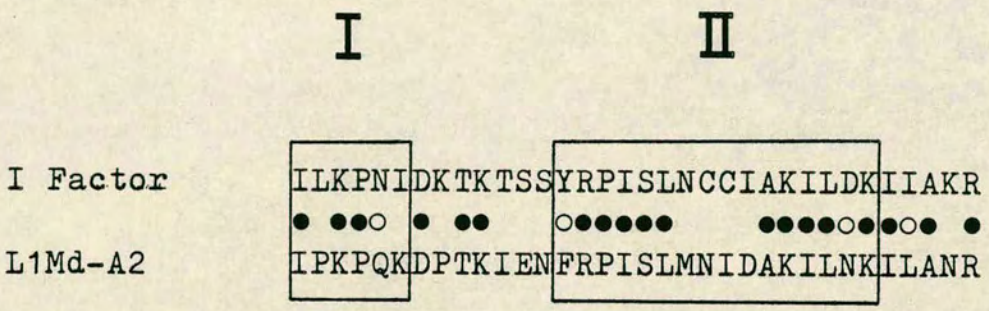

I
II
I Factor
ILKPNIDKTKTSSYRPISLNCCIAKILDKIIAKR
L1Md-A2
IPKPQKDPTKIENFRPISLMNIDAKILNKILANR

Figure 3.9B

suggesting that the putative reverse transcriptase gene of I is more closely related to that of LINEs than to retroviruses, introns, etc.

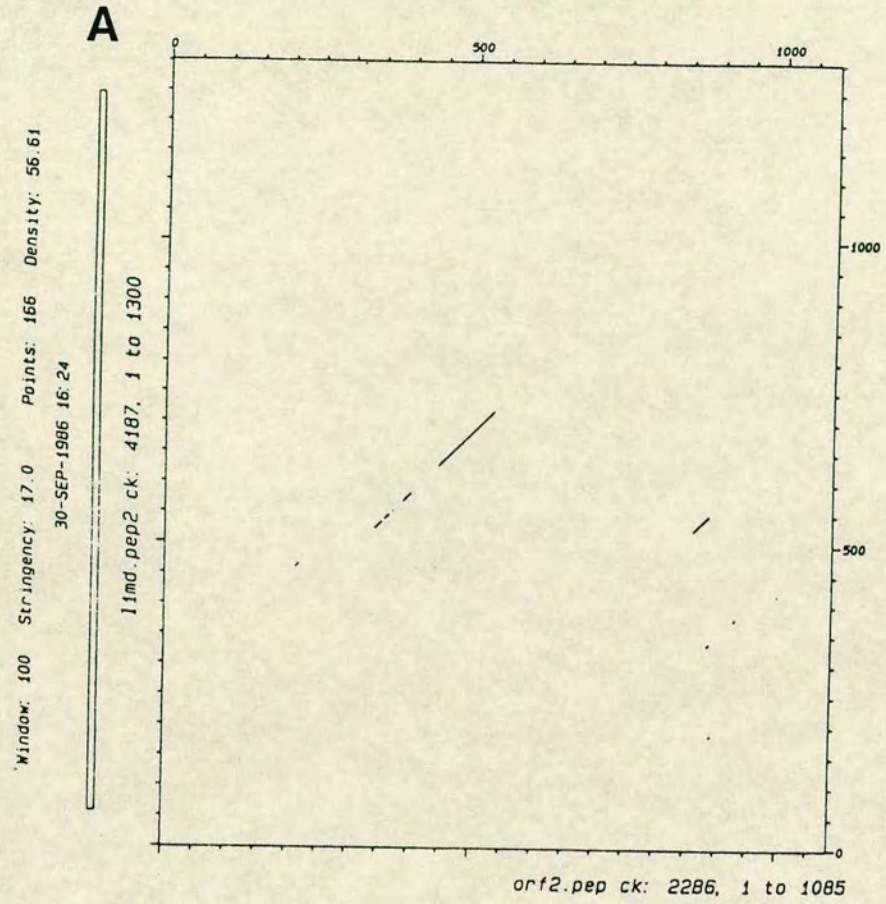
Burton et al. have compared cloned mouse and human LINE DNA by Southern blotting, and found two regions which cross hybridise most strongly. These are called CS1 and CS2, for "conserved sequence". Both are within the larger open reading frame of L1Md-A2. CS2 corresponds to the region of domain 2 of the reverse transcriptase, and hence the region most strongly conserved between LINE elements is also most strongly conserved in the I factor. This implies a constraint upon mutations of this region, suggesting that this region may be important for protein function. A comparison of the two protein sequences from the I factor and L1Md-A2 by a dot-plot shows a region of homology (Fig. 3.10A). This region corresponds to CS2. A similar comparison between I and RSV does not detect any homology. Clearly I is more closely related to LINEs than to retroviruses (Fig. 3.10B).

No homology can be found between ORF1 of I and the shorter open reading frame of L1Md-A2. Either the two protein sequences have diverged beyond the point where any similarity could be detected, or else they are completely unrelated. Within ORF1 of the I factor there is a sequence $CX_2CX_4HX_4C$, which is also found in basic nucleic acid binding proteins cleaved from the gag protein of retroviruses. It is thought that in viruses this protein interacts directly with the viral genomic RNA (Dickson et al., 1985; Copeland et al., 1983; Covey, 1986). It has been found in all retroviral gag proteins, CaMV viral coat protein, intracisternal A-type particles of Syrian Hamster and

Figure 3.10

A. Dot plot comparison of I factor ORF2 peptide sequence with L1Md ORF2 peptide sequence, using the method of Collins and Coulson (1986).

B. Dot plot comparison of I factor ORF2 peptide sequence and RSV reverse transcriptase peptide sequence. The parameters for this comparison were the same as those used in A.


A
Window: 100 Stringency: 17.0 Points: 166 Density: 56.61
30-SEP-1986 16:24
11md.pep2 ck: 4187, 1 to 1300
orf2.pep ck: 2286, 1 to 1085
0
500
1000
1000
500
0

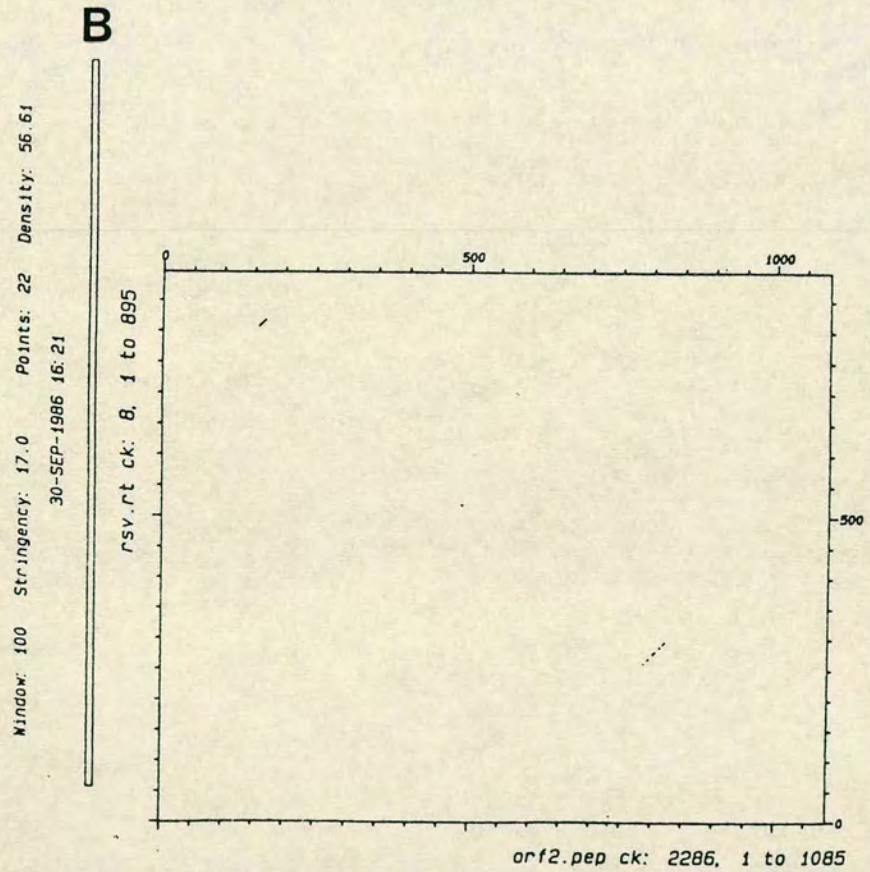

B
Window: 100 Stringency: 17.0 Points: 22 Density: 56.61
30-SEP-1986 16:21
rsv.rt ck: 8, 1 to 895
orf2.pep ck: 2286, 1 to 1085
0
500
1000
500
0

Figure 3.10

copia, but not in 17.6, the Ty element of S. cerevisiae or HBV (Covey, 1986). Several genes contain 2 copies of this sequence, the I factor contains 1 copy plus one imperfect copy (see Fig. 3.11). A protein which binds nucleic acid may play a role in I factor transposition or regulation. It is thought that two functions are encoded by I from genetic studies (Bregliano & Kidwell, 1983), namely a transposase and a regulator of transposition. It is possible that the transposase is a reverse transcriptase encoded by ORF2, and the regulator a nucleic acid binding protein encoded by ORF1. This will be discussed further in Chapter 5.

Figure 3.11

Comparison of ORF1 of the I factor with the conserved domains in viral nucleic acid binding proteins. Viral sequences are as follows: RSV, the gag gene of RSV; HTLV, the gag gene of HTLVI; CaMV, the coat protein of CaMV. The numbers indicate the numbers of amino acids from the start of the polypeptide or between conserved domains in the same polypeptide. Conserved amino acids are boxed.

From Fawcett et al. (1986).

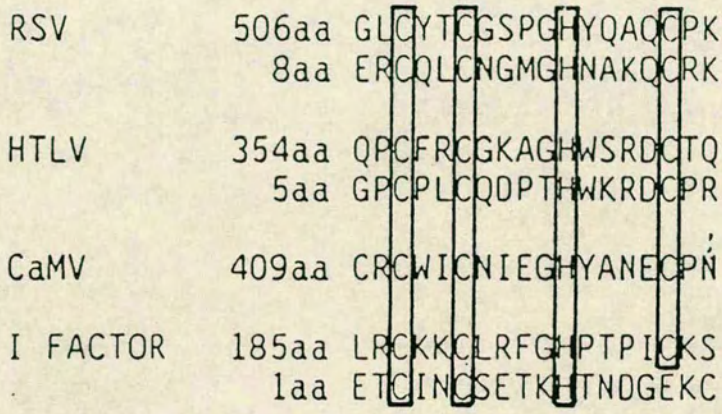

RSV
506aa GLCYTCGSPGHYQAQCPK
8aa ERCQLCNGMGHNAKQCRK
HTLV
354aa QPCFRCGKAGHWSRDCTQ
5aa GPCPLCQDPTHWKRDCPR
CaMV
409aa CRCWICNIEGHYANECPN
I FACTOR
185aa LRCKKCLRFGHPTPICKS
1aa ETCINCSETKHTNDGEKC

Figure 3.11

CHAPTER 4

Sequence Analysis of Other I Factor-Induced Mutations

4.1 Introduction

Several mutations of the *D. melanogaster* *white* locus which arose following a number of I-R dysgenic crosses have been described by Bucheton *et al.* (1984) and Sang *et al.* (1984). Eight mutant strains were established, called w^{IR1-8}. These fell into two classes. w^{IR1-6} were associated with insertions of an apparently identical 5.4kb element - the putative I factor. The cloning and sequencing of the I factor from w^{IR1} has been described in Chapter 3. All of these insertion mutations retained some eye colour. Strains w^{IR1-6} have been tested for the ability of the I factor to induce dysgenesis in an IR dysgenic cross. All have been found to be active, with the exception of w^{IR2} (A. Pelisson, unpublished data). Mutant strains w^{IR7} and w^{IR8} have white eyes and are associated with deletions of part of the *white* gene. No I factor sequences could be detected in the *white* gene of these two strains (Sang *et al.*, 1984), and it was uncertain whether the deletions arose as a consequence of IR dysgenesis (perhaps due to aberrant insertion events) or whether they were random mutations picked up in the screening for IR-induced *white* mutations. The positions of the I factor insertions, and the deletion breakpoints, are shown in Figure 4.1.

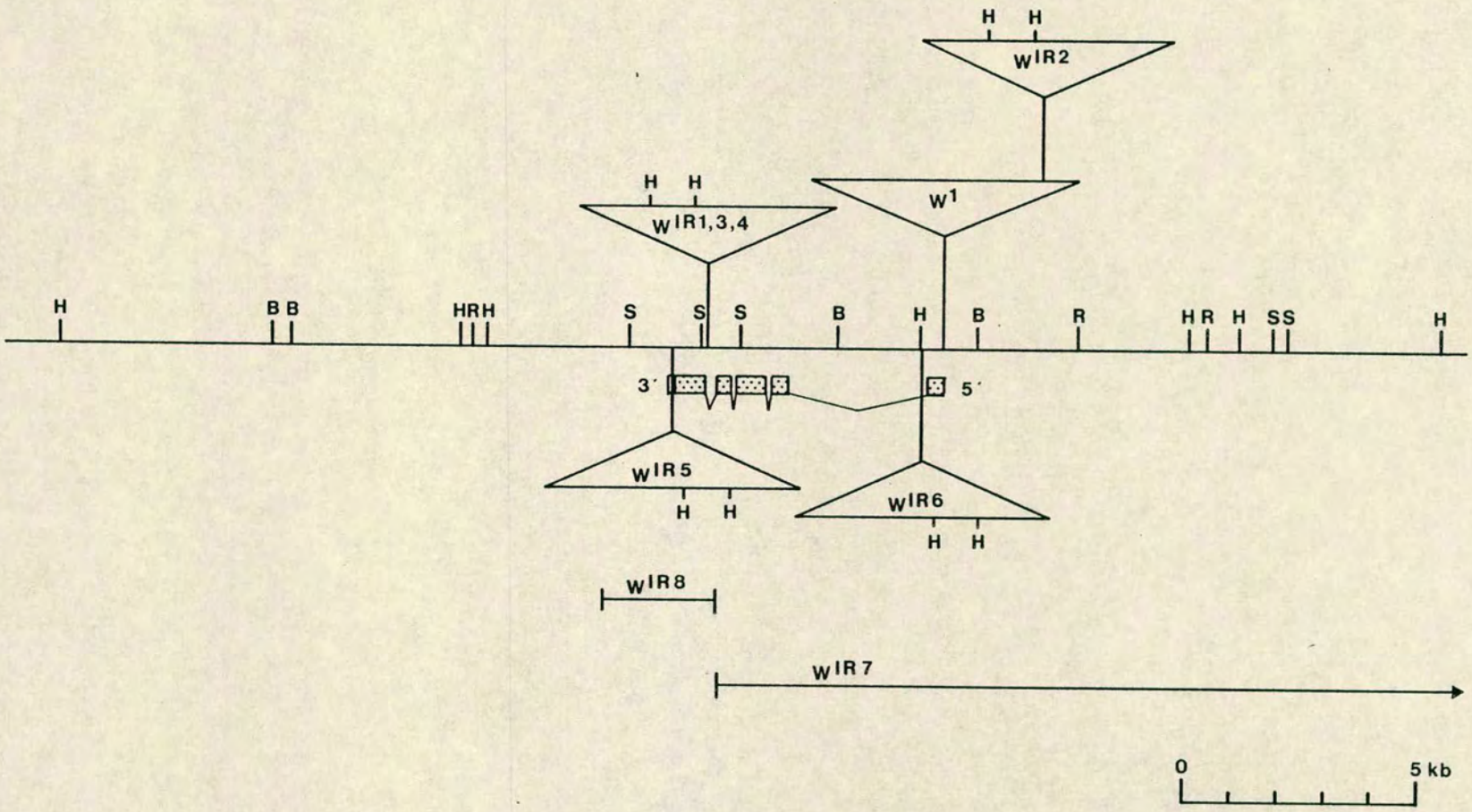
In this chapter the mapping of the insertion points will be described and comparisons will be made between the ends of the different elements and the target sites. This information was gained by cloning restriction fragments from either end of each element that

Figure 4.1

Restriction map of the white gene showing the insertion points of the I factors in strains w^{IR1-6} and the deletion breakpoints in strains w^{IR7} and w^{IR8}. Stippled boxes show the putative white exons.

Symbols:- H, HindIII; B, BamHI; R, EcoRI; S, SalI.

Figure 4.1


H H
wIR2
H H
wIR1,3,4
w1
H
B B
HRH
S
S
S
B
H
B
R
HR H SS
H
3'
5'
wIR5
H H
wIR6
H H
wIR8
wIR7
0
5 kb

spanned the junction between white and I factor sequences. Some possible explanations for the phenotypic effects of the insertions will then be proposed.

In addition to the insertion mutations the deletion mutations w^{IR7} and w^{IR8} have also been analysed in this project. The breakpoints have been cloned and sequenced to look for any evidence of I factor activity, such as I factor sequence or a sequence duplication.

Finally, the sequence of the ends of an I factor inserted into the bithorax complex will also be described. This spontaneous mutation of bithorax had been found by Peifer and Bender (1986) to be associated with an I factor insertion and the resulting strain was called bx^{F31}.

4.2 The w^{IR3} mutation

The position of the I factor insertion in w^{IR3} had previously been mapped, by Southern blotting, to a position very close to the w^{IR1} insertion point (Bucheton et al., 1984). The ends of the w^{IR3} I factor had been cloned on HindIII fragments containing I factor DNA and flanking white sequences. The resulting lambda clones (in vector λNM1149 were called λI451 (left end of the I factor) and λI452 (right end) (Fig. 4.2).

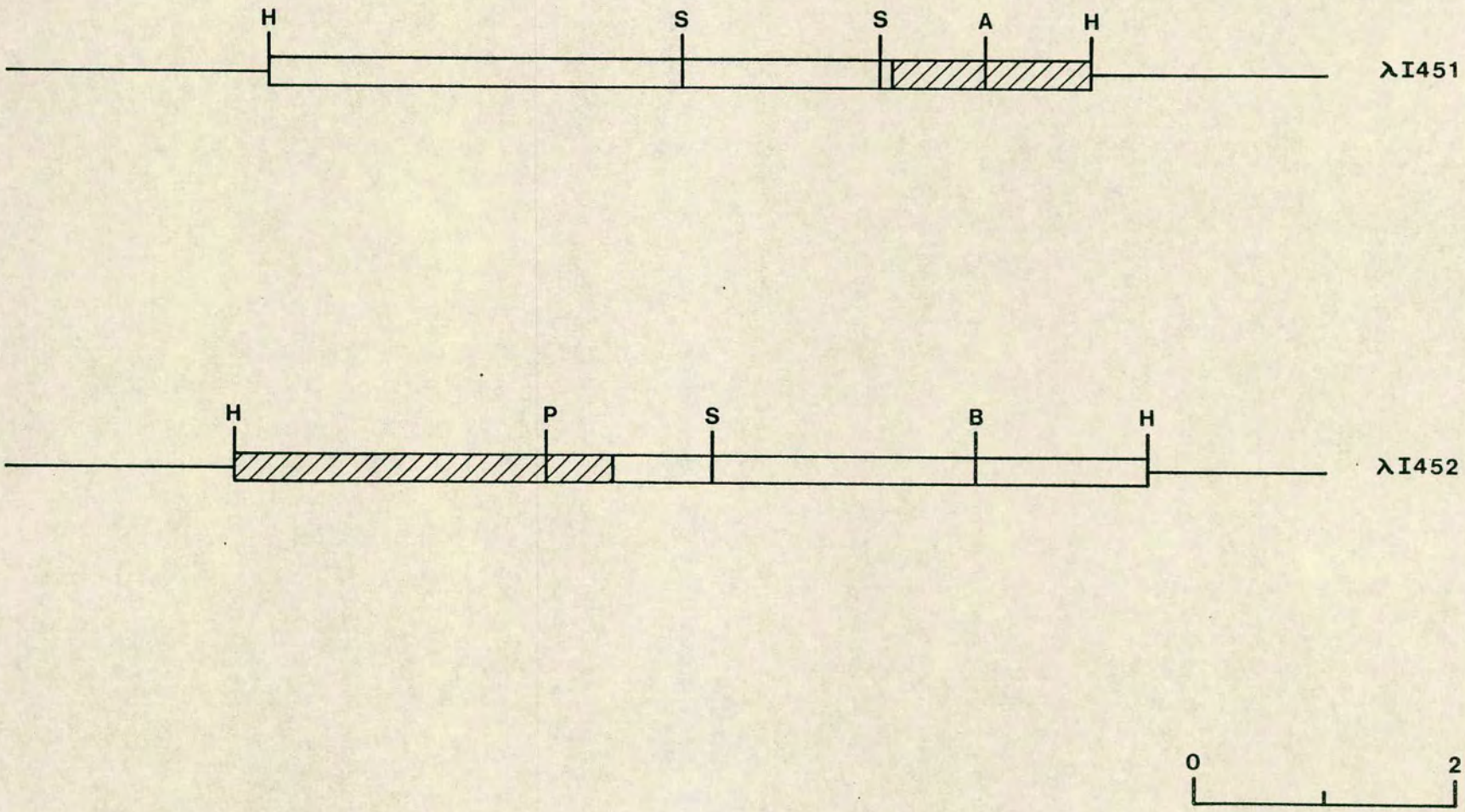
For this project, the two ends were subcloned into M13. λI451 was restricted with HindIII and SalI to generate a fragment equivalent to

Figure 4.2

Maps of the insertions in the lambda clones λI451 and λI452. Open boxes represent white sequences, cross-hatched boxes represent I factor sequences. Single lines represent lambda arms.

Symbols:- H, HindIII; S, SalI; A, AvaI; P, PstI; B, BamHI.

Figure 4.2



I770 from $\underline{w}^{IR1}$ (Fig. 3.2). Fragments were ligated into HindIII/SalI-cut mp18 and transformed into NM522. The resulting white plaques were screened with pCS155 (Fig. 3.1). Templates were prepared from positive plaques and were sequenced from the SalI site through flanking white DNA and over the end of the I factor. This enabled the insertion point to be mapped accurately, and it was found to be exactly as that of $\underline{w}^{IR1}$. One hundred and ten bases of I factor DNA were determined and found to be identical to the first 110 bases of the $\underline{w}^{IR1}$ I factor except for base number 3. In $\underline{w}^{IR1}$ this is a T, in $\underline{w}^{IR3}$ it is a G.

The right end of the $\underline{w}^{IR3}$ I factor was subcloned into M13 by restricting λI452 with HindIII and XbaI. There is an XbaI site in the $\underline{w}^{IR1}$ I factor at position 5108 (Fig. 3.5), which is close enough to the end of the element to sequence from it over the end of the element into flanking white sequence. Fragments were ligated into HindIII/XbaI-cut mp19, transformed into NM522 and the resulting white plaques screened with pCS155. Templates were made from positive plaques and were sequenced, from the XbaI site. The XbaI site is 263bp from the end of the I factor. A comparison between the corresponding $\underline{w}^{IR1}$ and $\underline{w}^{IR3}$ sequences showed them to be identical. The only (potential) difference found was in the number of TAA triplets at the end of the elements. $\underline{w}^{IR1}$ has 4 or 5, $\underline{w}^{IR3}$ has 5. The number of bases duplicated flanking the $\underline{w}^{IR3}$ I factor is 14, but it should be noted that this would correspond to a 13bp duplication in $\underline{w}^{IR1}$ as $\underline{w}^{IR3}$ contains an extra base (an A) within the duplication (Fig. 4.3). This extra base may have been present in the

Figure 4.3

Sequences at the ends of the I factor insertions. The sequences of the I factor ends are shown in upper case, flanking sequences are shown in lower case. Target site duplications are boxed. Dotted lines indicate bases which could be part of the duplication or part of the I factor. The tandem repeats adjacent to the left end of the w^{IR6} I factor are indicated.

From Fawcett et al. (1986).

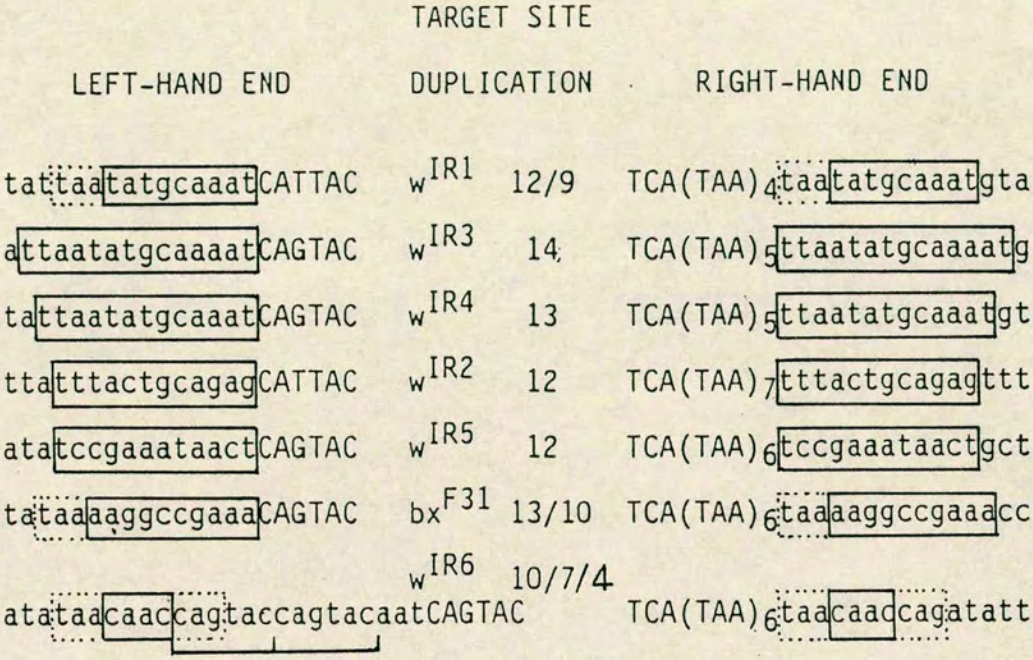

TARGET SITE
LEFT-HAND END
DUPLICATION
RIGHT-HAND END
tattaatatgcaaatCATTAC
wIR1
12/9
TCA(TAA)4taatatgcaaatgta
attaatatgcaaaatCAGTAC
wIR3
14
TCA(TAA)5ttaatatgcaaaatg
tattaatatgcaaatCAGTAC
wIR4
13
TCA(TAA)5ttaatatgcaaatgt
ttatttactgcagagCATTAC
wIR2
12
TCA(TAA)7tttactgcagagttt
atatccgaaataactCAGTAC
wIR5
12
TCA(TAA)6tccgaaataactgct
tataaaggccgaaaCAGTAC
bxF31
13/10
TCA(TAA)6taaaggccgaaacc
wIR6
10/7/4
atataacaaccagtaccagtacaatCAGTAC
TCA(TAA)6taacaacagatatt

Figure 4.3

target site of the parent or it may have arisen as a consequence of I factor insertion - the parental target site was not sequenced and hence the origin of the extra base is not known. Unlike $\underline{w}^{IR1}$ the mutation has not been assigned to the X chromosome of either the maternal ($\underline{seF}_8$) or paternal (Luminy) strain. However, if the extra base was present before I factor insertion then the parent chromosome must have come from Luminy as the sequence of $\underline{seF}_8$ is known in this region (from the $\underline{w}^{IR1}$ insertion).

4.3 The $\underline{w}^{IR4}$ mutation

The position of the $\underline{w}^{IR4}$ I factor had been mapped by Southern blotting to a position very close to the $\underline{w}^{IR1}$ and $\underline{w}^{IR3}$ insertion points and in the same orientation (Sang et al., 1984). This DNA had not previously been cloned.

For this project a genomic library of this strain was constructed in phage λNM1149. DNA was first extracted from flies (Chapter 2) and then restricted with HindIII. Fragments were ligated into HindIII-cut λNM1149 and the resulting recombinant molecules packaged in vitro (Chapter 2). The phage produced were plated on NM514 cells, which allow only recombinant λNM1149 phage to grow and hence remove the background of parental phage. This library should have contained phage carrying the left end and right end of the $\underline{w}^{IR4}$ I factor. These phage would correspond to the $\underline{w}^{IR3}$ clones λI451 and λI452. Approximately 50,000 plaques were screened with the white probe pI54 (Fig. 3.1) to detect phage carrying white DNA from the region of the I factor insertion. Eight potential positive plaques were picked up in

this first screen. As plaques on the plates were confluent, phage from the general area of the positive spot were picked into phage buffer and plated (on NM514 cells) at a dilution which gave well separated plaques on the plates. These plaques were screened again with pI54 and several positives from each plate picked to a grid. Duplicate filters were made from this grid.

The probe pI54 detected both phage carrying the left end and the right end of the I factor. In order to distinguish these, one filter made from the grid was probed with pI769 (white DNA to the left of the I factor insertion and hence the left end of I would be detected) and the other filter was probed with pI768 (white DNA to the right of the insertion) (see Fig. 3.1). Two spots hybridising with pI769 and two spots hybridising with pI768 were picked. Plate lysates, and then liquid lysates, were made from each phage (using ED8654 as host rather than NM514 as ED8654 is a more healthy strain and hence gives better yields of phage. ED8654 allows non-recombinants as well as recombinants to grow and hence was not used for λNM1149 library construction).

DNA was prepared from these phage (see Chapter 2). In order to confirm that the inserts were correct the phage DNAs were restricted with HindIII. The digests were compared, by agarose gel electrophoresis, with HindIII-cut λI451 and λI452 (see Fig. 4.4A). Phage containing inserts identical to λI451 contained the left end of the w^{IR4} I factor, and were called λI421. Phage containing inserts identical to λI452 contained the right end of the w^{IR4} I

Figure 4.4

A. Restriction of clones λI421 and λI422. Tracks were loaded as follows:-

1. λcI857 HindIII markers; 2. λI451 uncut; 3. λI451 HindIII; 4. λI452 uncut; 5. λI452 HindIII; 6. λI421 uncut; 7. λI421 HindIII; 8. λI422 uncut; 9. λI422 HindIII; tracks 10-13 as 6-9 respectively (identical clones). The 6.9 kb and 6.2 kb bands, corresponding to the left and right ends of the I factor respectively, are indicated.

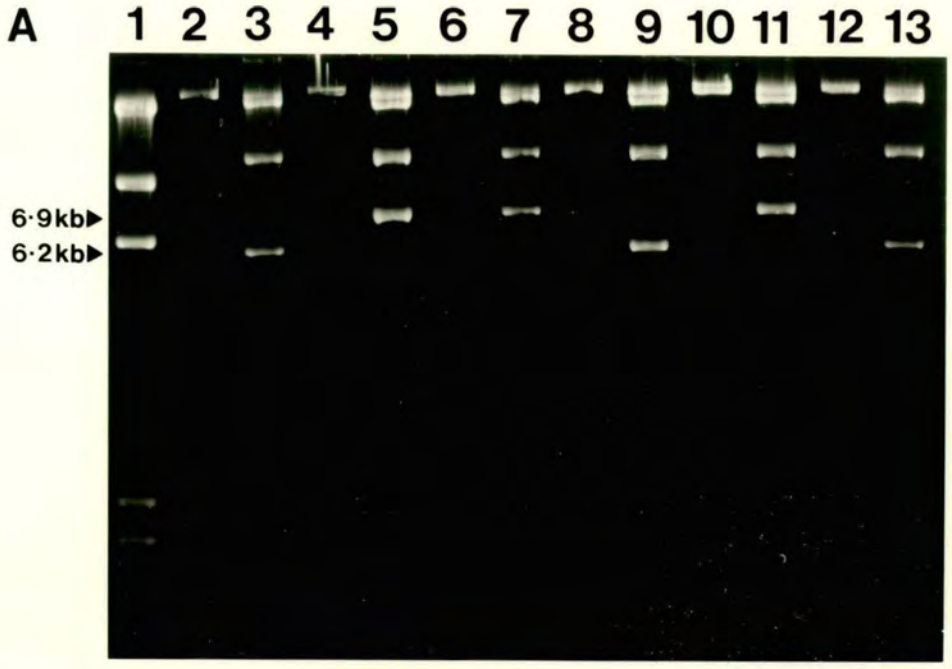
B. Restriction of λI423. Tracks loaded as follows:-

1. λI423 uncut; 2. λI423 HindIII; 3. λCI857 HindIII markers. The 7.1 kb band, corresponding to the I factor left hand end, is indicated.

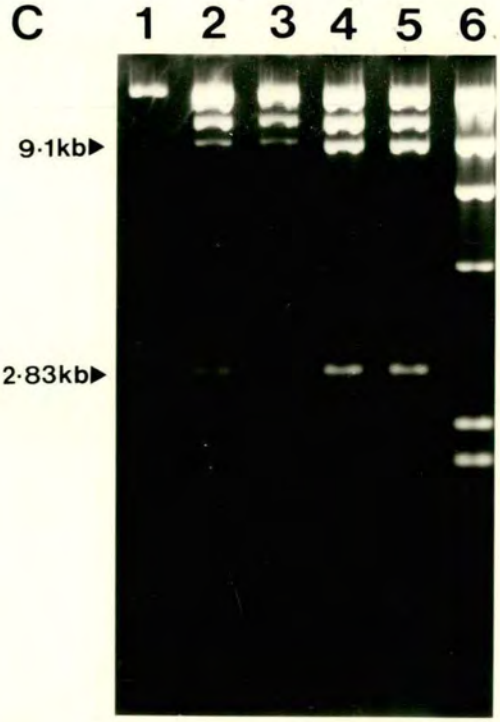
C. Restriction of λI424. Tracks loaded as follows:-

1. λI424 uncut; 2-5. λI424 HindIII (four identical clones); 6. λCI857 HindIII. The 9.1 kb and 2.83 kb bands of the insert in phage λI424 are indicated.

Figure 4.4


A
1 2 3 4 5 6 7 8 9 10 11 12 13
6·9kb▶
6·2kb▶


B
1 2 3
7·1kb▶


C
1 2 3 4 5 6
9·1kb▶
2·83kb▶

factor and were called λI422. These are shown in Figure 4.5.

Fragments containing the junction between the ends of the I factor and white DNA were subcloned into M13 in the same way as for w^{IR3}. λI421 was restricted with HindIII and SalI, ligated into HindIII/SalI-cut mp18 and transformed into NM522. White plaques were screened with pCS155 and positives picked and sequenced (from the SalI site). λI422 was restricted with HindIII and XbaI, ligated into HindIII/XbaI-cut mp19 and transformed into NM522. Plaques were screened with pCS155 and positives picked and sequenced (from the XbaI site).

The w^{IR4} I factor has inserted into the same position as the w^{IR1} and w^{IR3} I factors, indicating that this is an insertion hotspot. One hundred and fifteen bp of sequence was determined at the left end of the I factor, this is identical to the w^{IR1} I factor sequence except at position 3 which, like w^{IR3}, is a G not a T. The 263bp from the XbaI site to the right end of the I factor are identical to those of w^{IR1}. There are 5 TAA triplets at the end of the element, and a target site duplication of 13 bases. The duplication starts at the same base as the duplication in w^{IR3}, but is one base shorter because of the extra base (A) in w^{IR3} (Fig. 4.3).

4.4 The w^{IR5} mutation

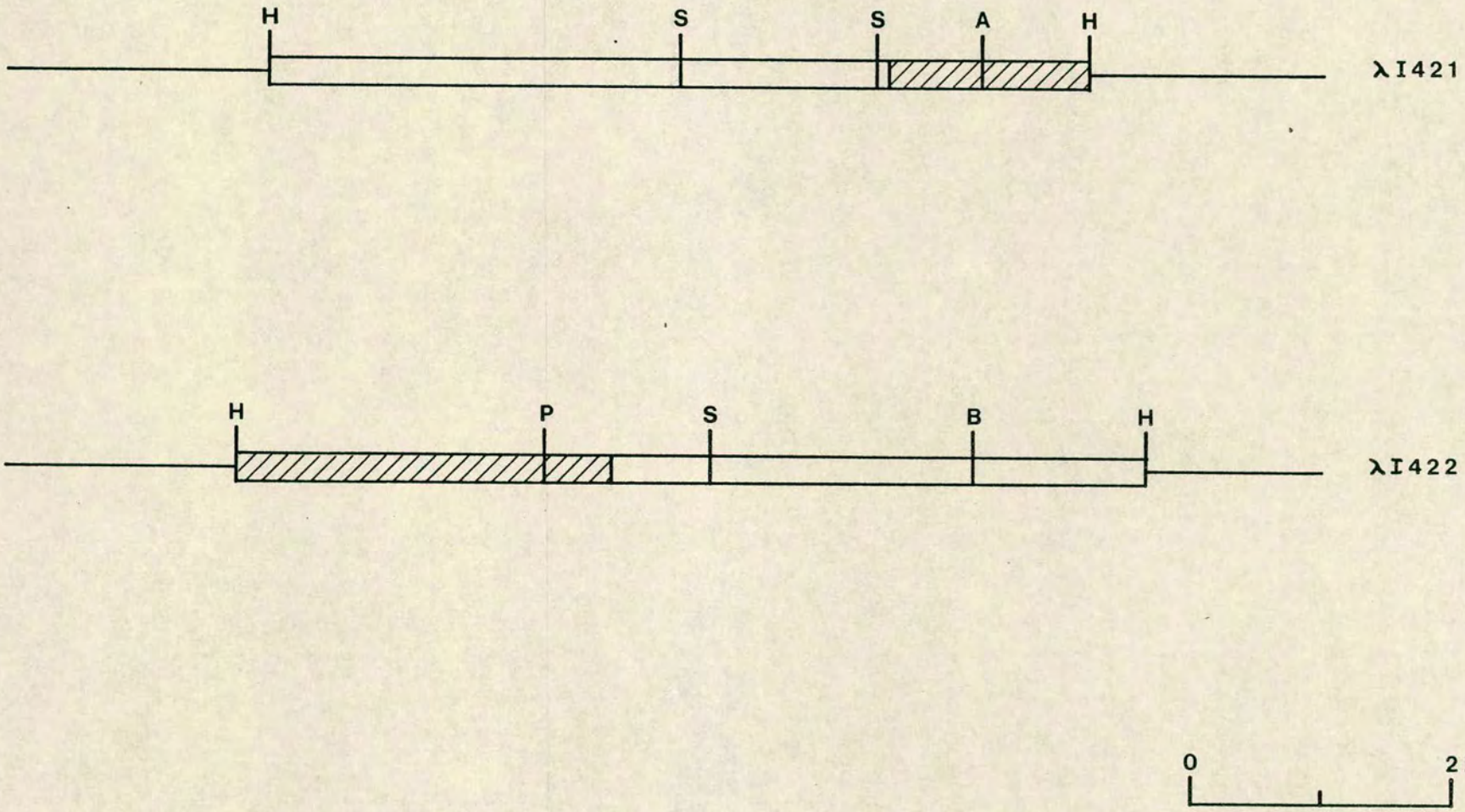
Sang et al. (1984) had mapped the insertion associated with the w^{IR5} mutation by Southern blotting. It mapped to the 3' end of the gene, possibly just within or just beyond the coding region. The orientation with respect to the white gene was found to be opposite

Figure 4.5

Restriction maps of the insertions in clones λI421 and λI422. Open boxes represent white sequence, cross-hatched boxes represent I factor. Single lines represent lambda arms.

Symbols:- H, HindIII; S, SalI; A, AvaI: P, PstI; B, BamHI.

Figure 4.5



that of $\underline{w}^{IR1,3 \text{ and } 4}$. The left and right ends of this element had already been cloned, on HindIII fragments, in λNM1149 (Sang et al., 1984) and had been subcloned on SalI/HindIII fragments, into pUC8. pI60 is the right end of the I factor, pI61 is the left end (Fig. 4.6).

For this project the ends were subcloned from pUC8 into M13. To subclone the right hand end pI60 was restricted with SalI and XbaI and the fragments subcloned into SalI/XbaI-cut mp19. White plaques produced upon transformation of NM522 were screened with pCS156 (Fig. 3.1) and positives picked, templates made and sequenced from the XbaI site.

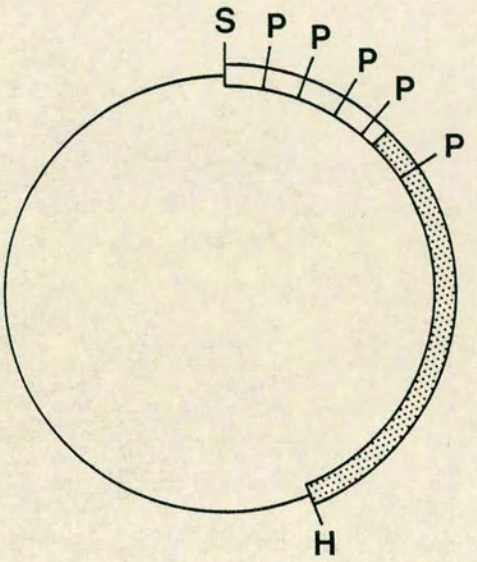
The left end of the I factor was subcloned in the following way. pI61 was cut with HindIII and SalI to excise the insert, and the whole digest run on an agarose gel. The insert band was cut out and the DNA recovered (see Chapter 2). The DNA was restricted with HpaII. There is a HpaII site in the I factor, 160bp from the left hand end (see Fig. 3.5), which is suitable for sequencing over the I/white junction. The resulting fragments were end-repaired (see Chapter 2) and ligated into HincII-cut mp19. This was transformed into NM522 and resulting white plaques screened with pI770. Positives were picked, templates made and sequenced. As the fragments were blunt ended it was not possible to select the orientation of cloning. Clones containing the desired fragment in the correct orientation (i.e. sequenced from the HpaII site across the end of the I factor) were identified by comparison of the sequence with that of the $\underline{w}^{IR1}$ I factor.

Figure 4.6

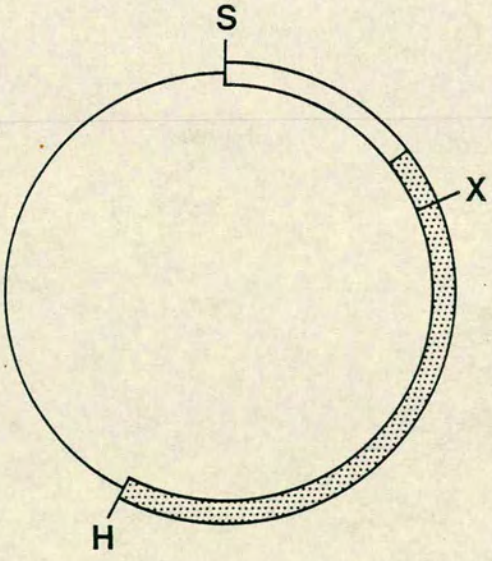
Plasmids pI61 (left end of w^{IR5} I factor) and pI60 (right end of w^{IR5} I factor). Open boxes represent white DNA, stippled boxes represent I factor DNA and single line represents pUC8 DNA.

Symbols:- S, SalI; P, HpaII; H, HindIII; X, XbaI.

Figure 4.6



pI61



pI60

The $\underline{w}^{IR5}$ I factor has 6 TAA triplets at the right hand end, and a target site duplication of 12bp (Fig. 4.3). One hundred and sixty bp of sequence was determined for the left end of this I factor. Again the third base differs from that of $\underline{w}^{IR1}$ (G rather than T). There is an additional change, at base 125. In $\underline{w}^{IR1}$ this is A, in $\underline{w}^{IR5}$ it is C.

At the right hand end 263 bases have been determined, and the sequence is identical to that of $\underline{w}^{IR1}$.

4.5 The $\underline{w}^{IR2}$ mutation

The $\underline{w}^{IR2}$ mutation, like $\underline{w}^{IR1}$, arose from a cross between an $\underline{seF}_8$ female and a $\underline{w}^{1}$ $\underline{ct}$ $\underline{f}$ male. Unlike $\underline{w}^{IR1}$ however, the I factor inserted into the X chromosome from the male parent. The white gene on this chromosome already contained an insertion, known as the $\underline{w}^{1}$ mutation. This insertion is a 5.7kb F-like element (O'Hare et al., 1983), and the resulting phenotype is white eyes. The element had been mapped to the 5' end of the white gene, within a probable 5' leader sequence (O'Hare et al., 1984). It is in the same orientation, relative to the white gene, as $\underline{w}^{IR1}$ (Sang et al., 1984). The I factor in $\underline{w}^{IR2}$ has inserted into the F-like element (Fig. 4.1). Some eye colour is restored in $\underline{w}^{IR2}$. Two other insertions into the $\underline{w}^{1}$ element are also associated with partial restoration of eye colour. These are the $\underline{w}^{e}$ (white - eosin) and $\underline{w}^{h}$ (white - honey) insertions (Lindsley & Grell, 1968; O'Hare et al., 1983, 1984 and unpublished data).

The $\underline{w}^{IR2}$ I factor had already been cloned intact on a BamHI fragment to form the lambda clone λI531 (Sang et al., 1984). This clone contains the I factor, the entire $\underline{w}^{1}$ element and some flanking white DNA (Fig. 4.7).

For this project both ends of this I factor were subcloned into M13. The right hand end was subcloned by restricting λI531 with BamHI and XbaI and ligating the fragments with BamHI/XbaI-cut mp18. This was transformed into NM522 and white plaques picked to a grid. These were screened with pI52 (Fig. 3.1), positives were picked and templates made and sequenced from the XbaI site. Two hundred and sixty-three bp were determined and found to be identical to the sequence of $\underline{w}^{IR1}$. Seven TAA triplets were found at the end of this I factor.

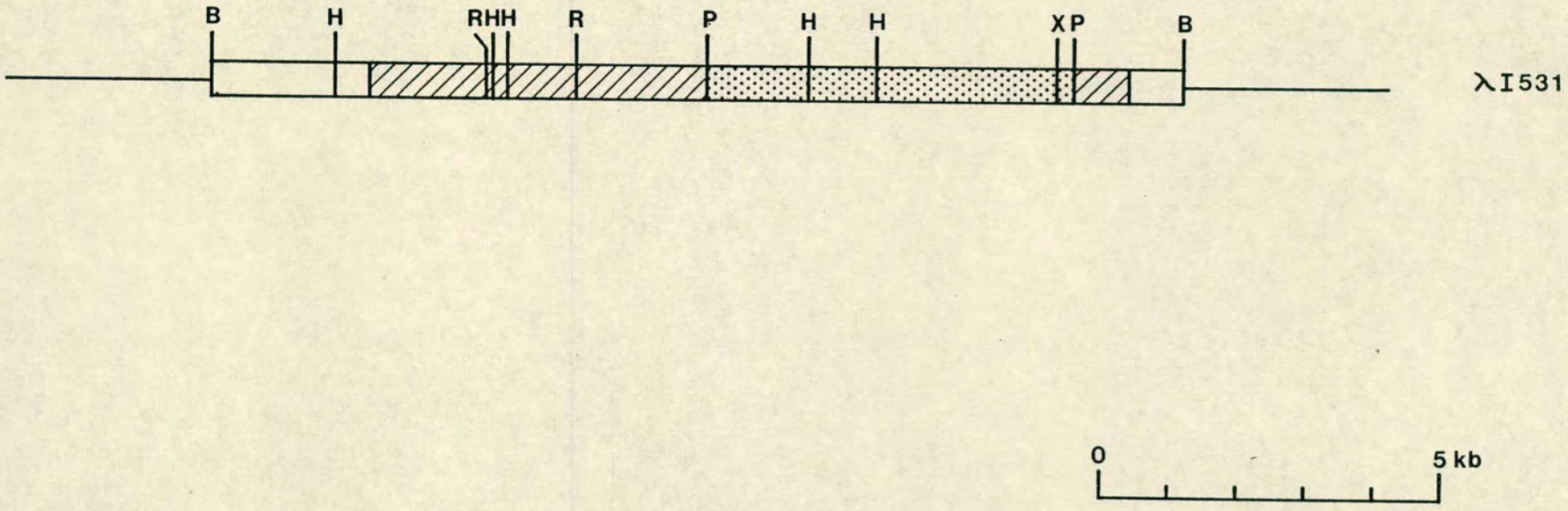
The left end of the I factor was cloned in two ways. Firstly, from the sequence of the right hand end it could be seen that the I factor had inserted adjacent to a PstI site which, because of its proximity to the end of the insertion, was included in the duplication of target sequence. λI531 was cut with HindIII and PstI, and the fragments obtained ligated with HindIII/PstI-cut mp19. Following transformation of NM522 white plaques were picked and screened with pI770. Positives were picked and sequenced from the PstI site. This showed that there is indeed a PstI site within the duplication, 2bp removed from the left end of the I factor. One hundred and seventy bp of sequence was determined of the left end, which was identical to the $\underline{w}^{IR1}$ sequence. In this element the third base is a T, as in $\underline{w}^{IR1}$, not G.

Figure 4.7

Restriction map of the insert in clone λI531, derived from $\underline{w}^{IR2}$ DNA. Stippled box represents I factor DNA, cross-hatched boxes represent $\underline{w}^{1}$ insert DNA and open boxes represent white DNA. Single line represents lambda arms.

Symbols:- B, BamHI; H, HindIII; R, EcoRI; P, PstI; X, XbaI.

Figure 4.7



In order to sequence more of the flanking DNA at the left end and thus to get the complete target site duplication, the left end was recloned from λI531 using EcoRI and HindIII. The fragments were ligated into EcoRI/HindIII-cut mp18, transformed into NM522 and white plaques subsequently screened with pI770. Positives were picked and templates made, but instead of sequencing them using the M13-specific sequencing primer the primer DF4 was used (see Chapter 2). This primer was synthesised specifically for sequencing the left end of the I factor, and is of the same sequence as bases 62 (5') - 48 (3') on the bottom strand. The information gained from this showed the target site duplication to be 12bp (Fig. 4.3).

The empty site from the w^1 element was also sequenced. The plasmid clone pw^1 (Fig. 4.8) was obtained from Kevin O'Hare. pw^1 was restricted with PvuII and the fragments ligated into SmaI-cut mp18. This was transformed into NM522 and white plaques picked to a grid. Duplicate filters were made, one was screened with pw^1 itself and the other with pAT153. Plaques positive only with pw^1 were picked and sequenced. The insertion point of the I factor was identified and it was confirmed that in this region the only rearrangement of w^1 DNA to have occurred as a consequence of I factor insertion was the duplication of 12bp flanking the element.

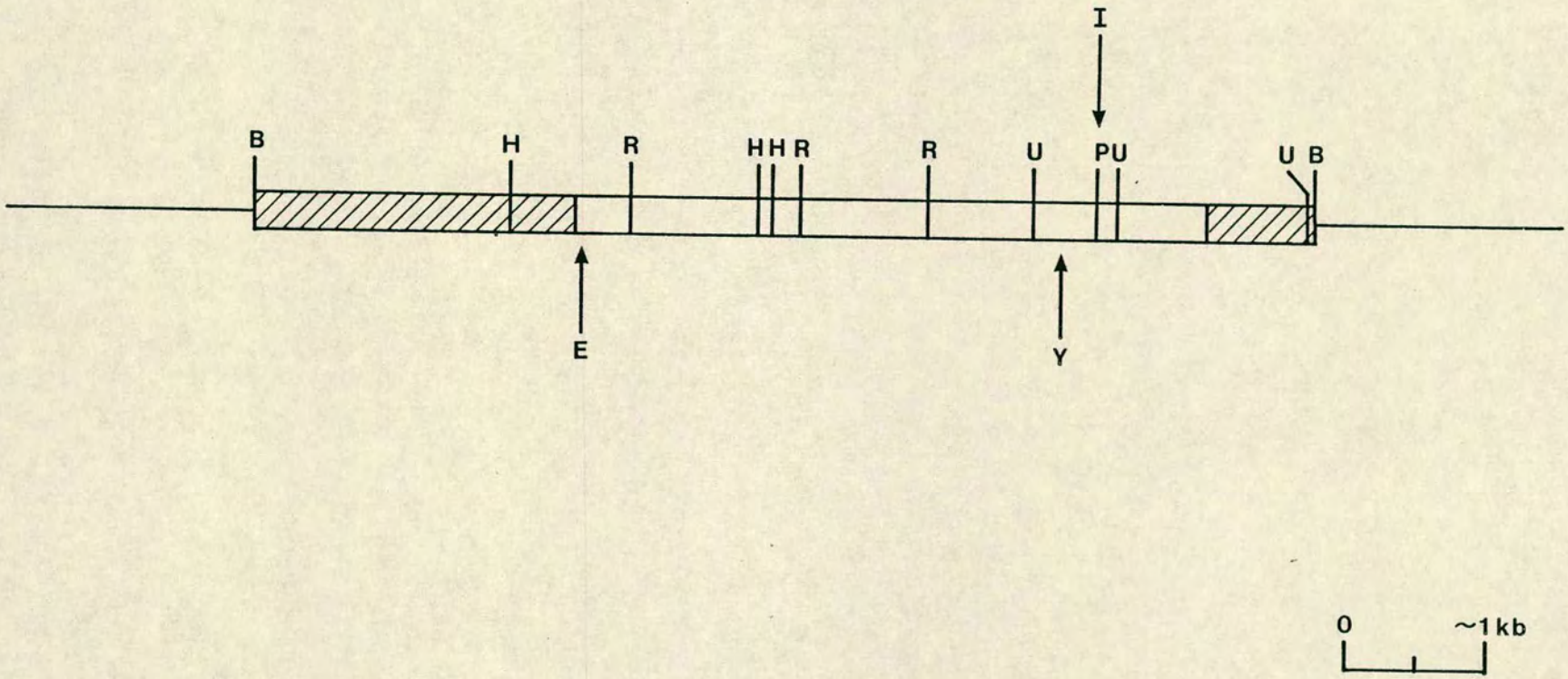
The reason for the inactivity of this element (with respect to causing dysgenesis) is unknown. Clearly there are no major deletions as the element measures 5.4kb and the ends are intact. Inactivity

Figure 4.8

Restriction map of the insert in plasmid clone *pw*^{1} (not drawn accurately to scale). Open box represents *w*^{1} (an F-like transposable element), cross-hatched boxes represent *white* DNA, single line represents plasmid vector DNA. The symbols I, E and Y show the insertion points of second transposable elements in the strains *w*^{IR2}, *white*-eosin and *white*-honey respectively.

Other symbols:- B, *Bam*HI; H, *Hind*III; R, *Eco*RI; U, *Pvu*II; P, *Pst*I.

Figure 4.8



could be caused by a minor mutation, perhaps within one of the open reading frames, or could be due to the site of insertion, being affected by expression of flanking sequences (a "position effect"). For example, I factor transcripts could be inactivated by hybridisation to an "anti-sense" RNA read from the white promoter and terminating within, or beyond, the I factor. This is illustrated in Figure 4.9.

4.6 The w^{IR6} mutation

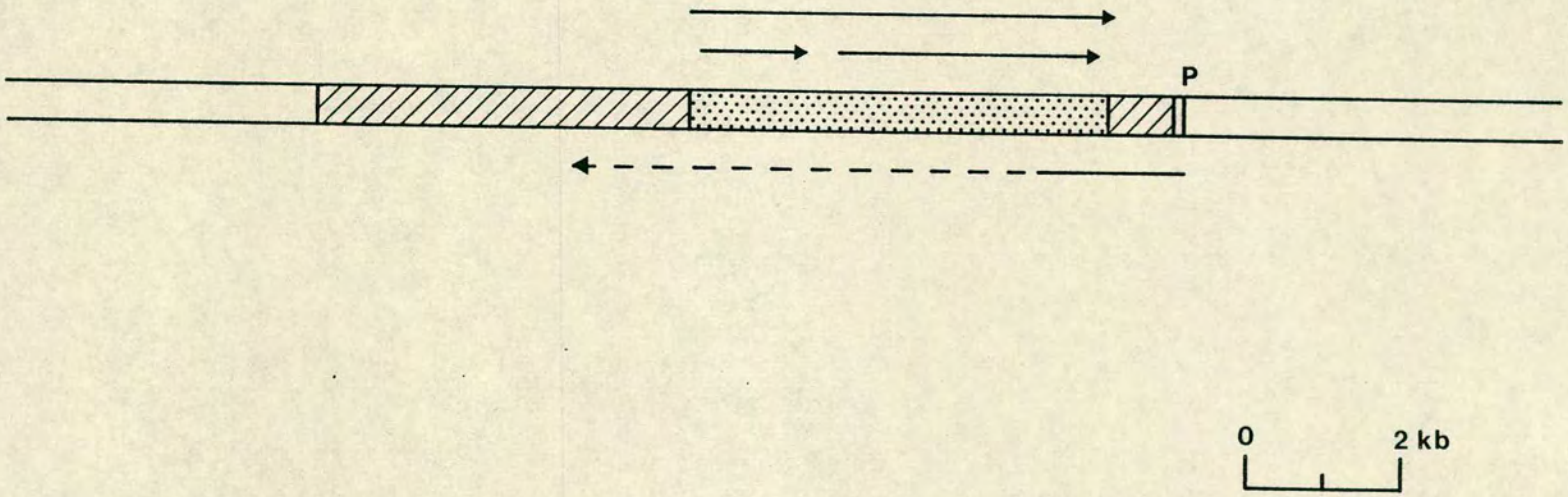
The w^{IR6} mutation had been mapped to a position just within the large intron of the white gene and was found to be in the opposite orientation to w^{IR1} (Sang et al., 1984). This I factor had not previously been cloned. It was known that the element had inserted very close to a HindIII site in the white DNA (Sang et al., 1984; Fig. 4.1), but it was not known whether there would be sufficient white DNA between the HindIII site and the right end of I to be detected by probing a HindIII library of w^{IR6} DNA.

For this project, the first library to be constructed was a HindIII library, in the hope that both ends of the I factor could be detected. Genomic DNA was extracted from w^{IR6} flies, and restricted with HindIII. The fragments were ligated into HindIII-cut λNM1149 and the resulting DNA molecules packaged in vitro (see Chapter 2). The library of phage produced was plated on NM514 and approximately 50,000 plaques were screened with pI52 (Fig. 3.1). Seven potential positive plaques were found, these were picked and replated (on NM514) for single plaques. These were rescreened with pI52, and four of the

Figure 4.9

Possible inactivation of the w^{IR2} I factor by a position effect. Stippled box represents I factor, cross hatched boxes represent w^{1} insert DNA and open boxes represent white DNA. Arrows indicate potential transcripts from I factor promoters and the white promoter (P).

Figure 4.9



original seven were positive again (the remaining three must have been spots of non-specific hybridisation to the filters). Several positive plaques from each plate were picked to a grid, and screened with pI770 (for the left end of the element) and pI771 (for the right end). All the plaques were positive only with pI770. This suggested that the I factor had inserted too close to the HindIII site in the white gene for clones containing the right hand end of the I factor to be detected using a white probe.

Two of the clones which were positive with pI770 were picked and plate lysates made (Chapter 2). Liquid lysates were subsequently made from these (Chapter 2), and the DNA isolated from the phage. The DNA of one clone (λI423, Fig. 4.10) was restricted with HindIII to check that it contained a band of the correct size (about 7.1kb, see Fig. 4.4B).

To subclone the left end of the I factor into M13, this lambda clone (λI423) was restricted with HindIII and the insert band purified by preparative agarose gel electrophoresis. This fragment was subsequently cut with HpaII, the ends repaired and the fragments cloned into SmaI-cut mp19. Templates were made of the white plaques recovered following transformation of NM522 and one M13 clone contained the HpaII fragment that spanned the left hand end of the I factor. The 160bp of I DNA from the end to the HpaII site were found to be identical to w^{IR1} except for the third base - G instead of T. Beyond the end of the I factor the sequence does not go straight into white sequence as expected, but instead there are what appear

Figure 4.10

Restriction maps of the inserts from clones λI423 and λI424 (left and right ends of the w^{IR6} I factor, respectively). Stippled boxes represent I factor DNA, open boxes represent white DNA. Single line represents lambda arms.

Symbols:- H, HindIII; A, AvaI; B, BamHI; R, EcoRI; S, SalI; P, PstI.

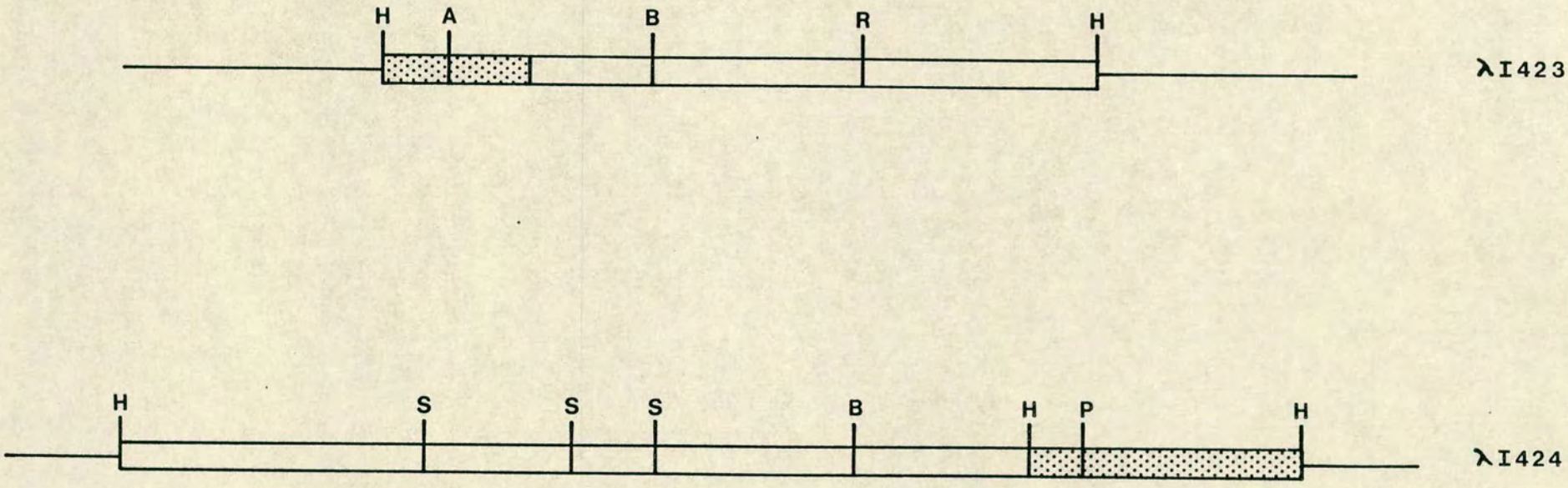


Figure 4.10

to be two tandem repeats of the last six bases of the I factor, separated from the end by an AAT triplet of unknown origin (Fig. 4.3). Beyond the second duplication the sequence changes to white sequence, but it is uncertain exactly where the white sequence starts as the first three bases of I (CAG) are also found in white at that position. This is reminiscent of the situation at the right end of the w^{IR1} I factor where the last TAA triplet could have come from the I factor or white. It is possible that the duplication of the end arose as a consequence of the integration mechanism, although one cannot rule out the possibility that the element already had this structure prior to integration.

The sequence of the white DNA was compared with the published sequence (O'Hare et al., 1984) and the I factor was found to have inserted 7bp from the HindIII site. This would result in a minimum of 13bp of white DNA beyond the right hand end of the I factor up to, and including, the HindIII site, possibly more depending upon the size of the duplication (if any). In theory this could have been enough to detect by plaque hybridisation but in practice the signal would probably have been too weak to detect.

To try and get the right hand end of the w^{IR6} I factor BamHI libraries were constructed. This involved cloning the I factor intact. Two vectors were used, EMBL4 (Murray, 1983) and λNM570-BV2 (Klein & Murray, 1979), but neither library contained the I factor despite several attempts (data not shown). This is perhaps not surprising as difficulty in cloning an intact I factor has been noted before

(Bucheton et al., 1984; Bender, unpublished data), although the reason is not understood. Sequences which contain inverted repeats (such as FB elements) or highly reiterated sequences are known to be unstable in phage lambda (Bellet et al., 1971; Paro et al., 1983), but the I factor contains neither of these.

The right hand end of the w^{IR6} I factor was finally cloned from a library of genomic DNA partially cut with HindIII. The fragment containing the right hand end of the element plus adjacent white DNA is approximately 12kb (see Fig. 4.1). This is too big to clone into λNM1149 (upper limit 11kb) so instead the vector λNM762 was used (see Chapter 2).

w^{IR6} DNA was digested with 0.5 and 0.25 units HindIII/ug DNA and aliquots removed after 3, 6, 9, 12 and 15 minutes. These aliquots were pooled to give a sample of DNA digested to varying extents. DNA digested with 0.25u/ug was less well cut. Both samples of DNA were ligated with HindIII-cut λNM762 and the resulting molecules packaged in vitro. λNM762 is a replacement vector and hence can re-clone its own central fragment as well as donor molecules. This central fragment contains an amber suppressor so that phage containing the fragment will generate blue plaques when plated on a host containing a lacZ amber mutation (in the presence of X-gal). The host strain used here was NM430 (Chapter 2). This therefore gives a visual assay for the frequency of recombinant formation, as recombinant phage give white plaques. The ratio of blue:white should be 1:1. Samples of both w^{IR6} libraries were plated first of all on NM430), and blue and

white plaques counted. The ratio of the 0.5u/ug library was 1:1, the 0.25u/ug library had rather more blue than white. This was not surprising however as the number of fragments put into the ligation would have been smaller as the DNA was less well digested.

The libraries were plated on ED8654. Approximately 50,000 plaques were screened, which corresponds to 25,000 recombinants for the 0.5u/ug library and less for the 0.25u/ug library. The probe used was pI54 (Fig. 3.1). Three positives were found, two from the 0.5u/ug library and one from the 0.25u/ug library. These were picked and plated to give single plaques. Duplicate filters were made from each plate, one was screened with pI54 and the other with pI905 (Fig. 3.2). pI905 is a plasmid containing the HindIII/PstI fragment from the right of the I factor. Only one plate contained plaques which were positive with both probes, these were derived from one of the phage from the 0.5u/ug library. Four well separated positive plaques were picked. Plate lysates and subsequently liquid lysates were made, and the DNA isolated.

To check the inserts of these four phages the DNA was restricted with HindIII. Two bands of insert DNA were expected, one of 9.1kb (white DNA) and one of 2.8kb (I factor DNA). These can be seen in Figure 4.4C. The clones were called λI424 (Fig. 4.10).

To clone the junction of the I factor/white DNA, λI424 was digested with PstI and BamHI, and the resulting fragments were ligated into PstI/BamHI-cut mp18. The ligated DNA was transformed into NM522

and white plaques picked and screened with pI771. Positives were picked, templates made and then sequenced not from the PstI site using an M13-specific primer but using primer DF3 (see Chapter 2). This primer was made specifically for sequencing the right hand ends of I factors, and is the same sequence as bases 5291 to 5305 of the w^{IR1} sequence (top strand). Fifty-nine bp of I factor sequence was obtained and found to be identical to the sequence of w^{IR1}. There are 6 or 7 TAA triplets at the end of the element, the uncertainty arising because of a TAA triplet in the white DNA as with w^{IR1} (Fig. 4.3). The number of bases duplicated flanking the I factor is 4, 7 or 10bp. A 4bp duplication would have arisen if the CAG triplet at the left end and the seventh TAA triplet at the right hand end were both I factor sequences; a 7bp duplication would arise if one of these triplets was white DNA and the other I factor DNA and a 10bp duplication would result if both triplets were white DNA. A 10bp duplication is more in keeping with the size of the other duplications found to date (9/12 - 14).

4.7 The bx^{F31} mutation

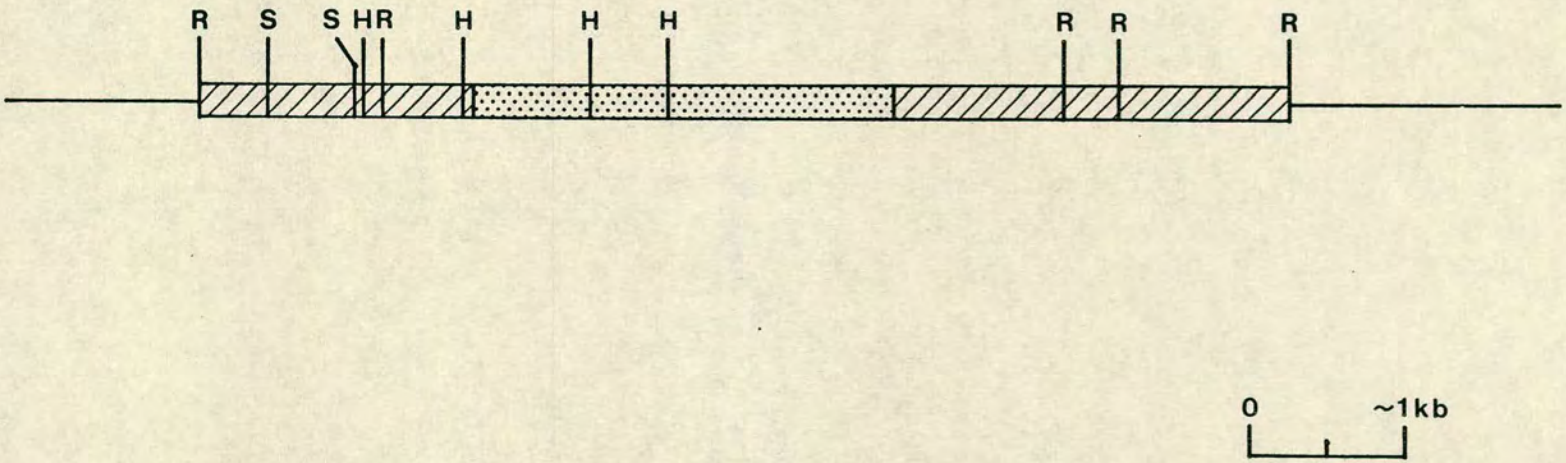
The I factor insertion associated with this mutation of the bithorax complex (Peifer & Bender, 1986) was detected as a spontaneous mutation, not the result of a dysgenic cross. This suggests that a low level of transposition of I factors must be possible in the inducer cytotype. This I factor had been cloned intact into a lambda vector (Peifer, unpublished data) and sent to this laboratory for analysis. A restriction map of the region flanking the insertion is shown in Figure 4.11. For this project, the ends of this I factor were

Figure 4.11

Restriction map of the insert in the lambda clone bx^{F31} (not drawn accurately to scale). Stippled box represents I factor DNA, cross hatched boxes represent bithorax DNA. Single line represents lambda arms.

Symbols:- R, EcoRI; S, SalI; H, HindIII.

Figure 4.11



subcloned into M13 for sequencing.

The I factor had inserted so that the left end is fairly close to a HindIII site. To clone this end the bx^{F31} clone was cut with HindIII and ligated into HindIII-cut mp18. White plaques found after transformation of NM522 were screened with pI770, and positives were picked and sequenced. Several clones were found whose sequence started in unrecognised DNA (bithorax) and then changed to I factor DNA. 70bp of I factor DNA was determined and found to be identical to the sequence of the w^{IR1} element.

The right hand end was cloned by digesting the bx^{F31} clone with HindIII and EcoRI and ligating the fragments into HindIII/EcoRI-cut mp18. This was transformed into NM522 and white plaques were screened with pI771. Positives were sequenced using the right end-specific primer DF3. Fifty-nine bp of I factor sequence was obtained and found to be identical with the w^{IR1} sequence. This element has 6/7 TAA triplets and a target site duplication of 10 or 13bp (Fig. 4.3).

4.8 The w^{IR7} mutation

The w^{IR7} mutation is a deletion mutation rather than an insertion. It is a large deletion which starts within the 0.86kb SalI fragment of the white gene and extends an unknown distance rightwards (Sang et al., 1984; Fig. 4.1). The roughest gene is probably included in the deletion however as w^{IR7} fails to complement roughest mutations. The DNA had not previously been

cloned and hence was done so as described below.

In order to select a suitable restriction enzyme with which to construct a genomic library, i.e. one which generated a fragment containing the breakpoint of a clonable size, $\underline{w}^{IR7}$ DNA was restricted with HindIII, EcoRI, SalI and BamHI. The digests were electrophoresed on an agarose gel and then transferred to nitrocellulose by the Southern blotting technique (Chapter 2). The filter was probed with pI54 (Fig. 3.1) and the result is shown in Figure 4.12B. EcoRI generated a fragment of the most suitable size (7.6kb). This digest is in fact a partial, the 7.6kb band is the lower of the two bands in the EcoRI track.

Genomic DNA was restricted with EcoRI and ligated with EcoRI-cut λNM1149. The resulting molecules were packaged in vitro and plated on NM514. Approximately 50,000 plaques were screened using pI769 as a probe, and five potential positives were found. These were picked, plated to give single plaques and rescreened with pI769. One plate had plaques hybridising with the probe. Two plaques were picked from this plate and plate lysates made from them. Liquid lysates were subsequently made and the DNA extracted from the phage. To check the inserts were of the right size these phage (called λI425 (Fig., 4.13)) were restricted with EcoRI and the fragments electrophoresed on an agarose gel. The result is shown in Figure 4.12C.

To subclone the deletion breakpoint into M13, λI425 was cut with EcoRI and SalI and ligated into EcoRI/SalI-cut mp18. Following

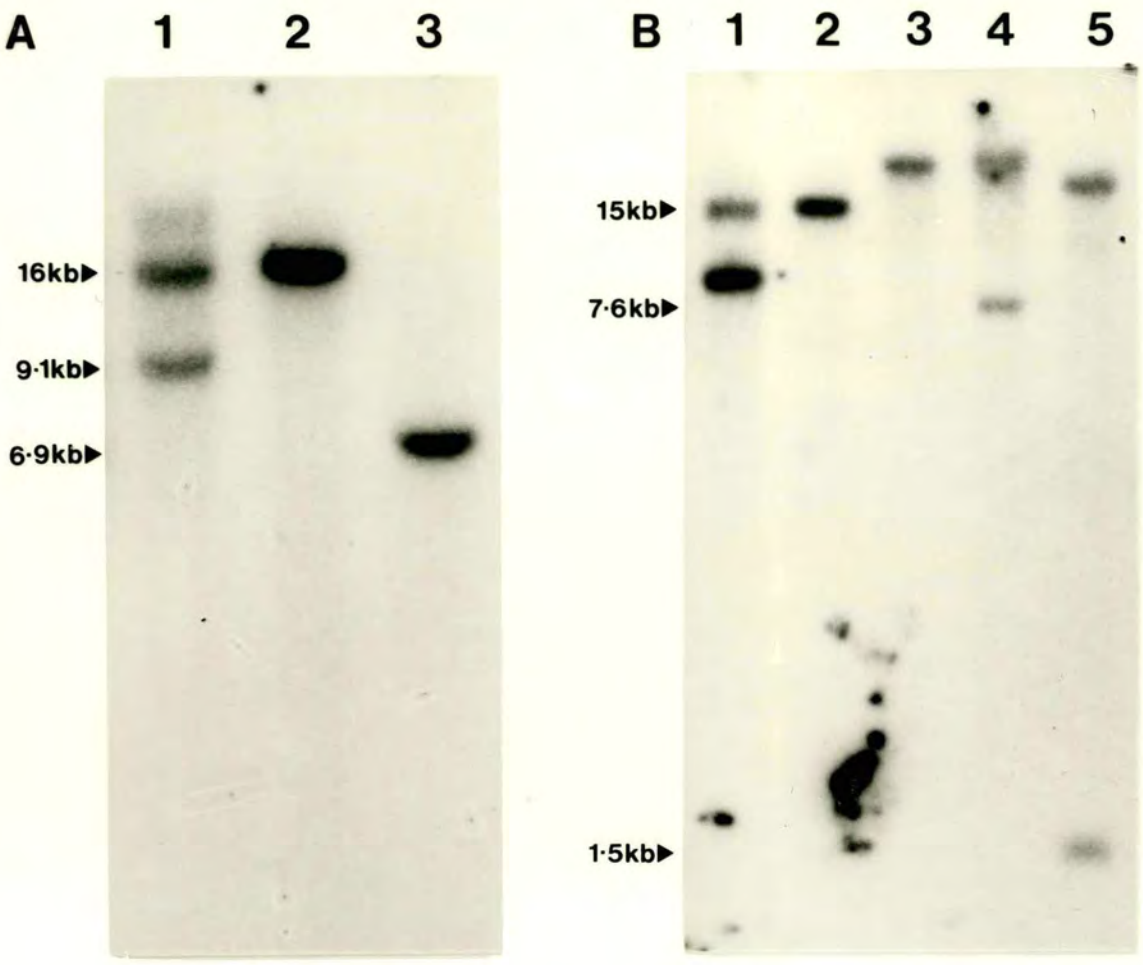
Figure 4.12

A. Southern blot of HindIII-cut Canton S DNA (track 1), w^{IR7} (track 2) and w^{IR8} (track 3). The filter was probed with pI769. The Canton S digest is partial, only the 9.1 kb band should be present. The 6.9 kb band in the w^{IR8} track is the fragment which contains the deletion breakpoint.

B. Southern blot of the following digests: Canton S cut with HindIII (track 1); w^{IR7} cut with HindIII, BamHI, EcoRI, SalI (tracks 2-5 respectively). The filter was probed with pI54. The Canton S digest and w^{IR7} (EcoRI) and w^{IR7} (SalI) digests are partials. The 7.6 kb EcoRI fragment (track 4) contains the w^{IR7} deletion breakpoint and is of a suitable size to clone into λNM1149.

C. Restriction digests of λI425 (w^{IR7}) and λI426 (w^{IR8}) to check insert sizes. Track 1, lambda HindIII markers. Tracks 2 and 3, λI425 cut with EcoRI. Tracks 4 and 5, λI426 cut with HindIII. Expected insert sizes of 7.6 kb (λI425) and 6.9 kb (λI426) are indicated. Track 5 shows the small insert in the second λI426 clone.

Figure 4.12



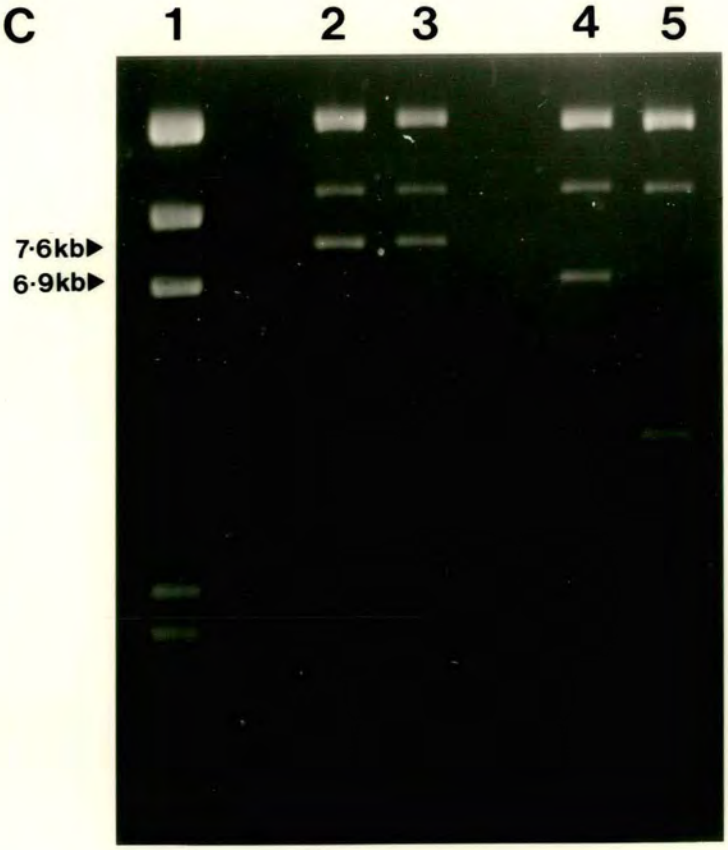
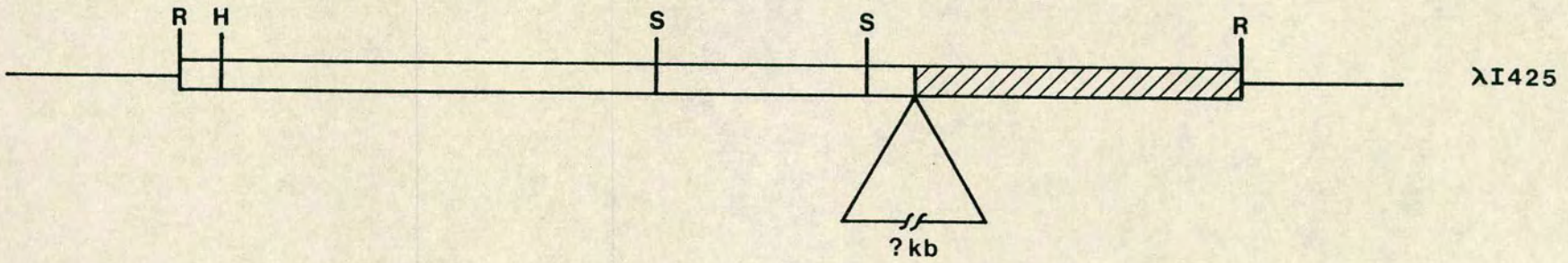


Figure 4.13

Map of clone λI425. Open box represents white DNA up to the deletion breakpoint. Cross-hatched box represents DNA on the other side of the breakpoint and is of unknown origin. Single line represents lambda arms.

Symbols:- R, EcoRI; S, SalI; H, HindIII.



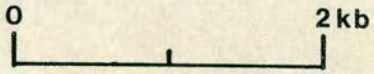


Figure 4.13

transformation of NM522 white plaques were picked and screened with pCS155. Positives were picked, templates made and sequenced from the SalI site. This is the SalI site which is 93bp to the left of the I factor insertion hotspot. By comparing the sequence of the w^{IR7} clone with that of the published white sequence (O'Hare et al, 1984) the breakpoint was identified 358bp to the right of the SalI site. The sequence of the breakpoint is shown in Figure 4.14. There is no evidence of the break having occurred as a result of I factor activity. There is no remnant of I factor sequence, no sequence duplication which may have been left behind after an abortive integration event, and the breakpoint does not correspond to the insertion hotspot. The possibility of I factor involvement cannot be ruled out, however.

4.9 The w^{IR8} mutation

The w^{IR8} deletion had been mapped by Sang et al. (1984). It appeared to have removed most of the CS155 and CS156 white DNA. Although the middle SalI site of the three in this region was thought to have been deleted the two outer sites were thought to remain, as the sizes of the adjacent SalI/HindIII fragments (I768 and I769) were the same as in a wild type strain. The deletion was measured at approximately 2.5kb.

In this project the size of the deletion was remeasured by digesting genomic w^{IR8} DNA with HindIII, running the fragments on an agarose gel and transferring the DNA to a nitrocellulose filter. The blot was probed with pI769, and the 9.1kb parental band (the lower

Figure 4.14

Sequences of the deletion breakpoints in $\underline{w}^{IR7}$, $\underline{w}^{IR8}$ and the $\underline{w}^{IR8}$-lambda deletion. All three are shown in relation to the wild-type white sequence. The breakpoints are indicated by arrows.

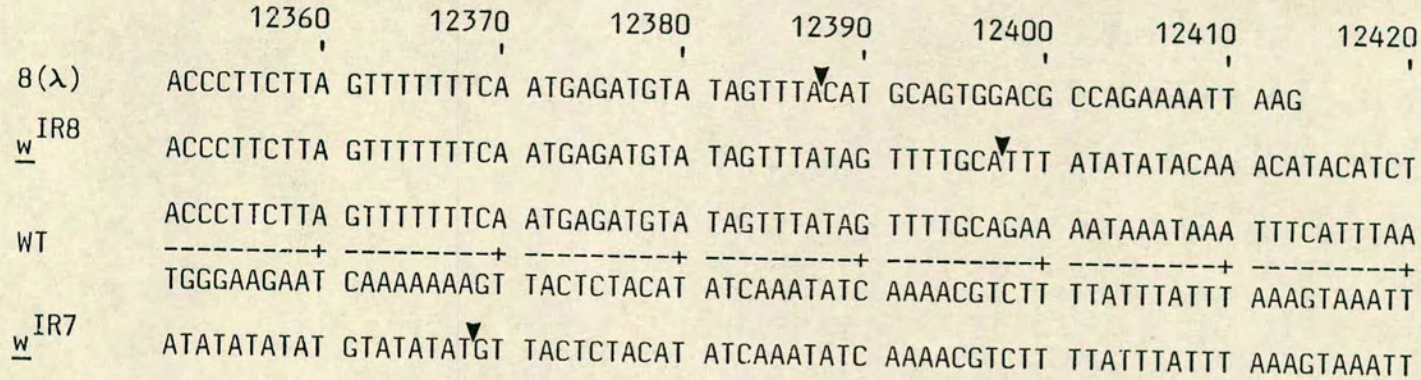

12360
12370
12380
12390
12400
12410
12420
8(λ) ACCCTTCTTA GTTTTTTTCA ATGAGATGTA TAGTTTACAT GCAGTGGACG CCAGAAAATT AAG
w IR8 ACCCTTCTTA GTTTTTTTCA ATGAGATGTA TAGTTTATAG TTTTGCATTT ATATATACAA ACATACATCT
WT ACCCTTCTTA GTTTTTTTCA ATGAGATGTA TAGTTTATAG TTTTGCAGAA AATAAATAAA TTTCATTTAA
TGGGAAGAAT CAAAAAAAGT TACTCTACAT ATCAAATATC AAAACGTCTT TTATTTATTT AAAGTAAATT
w IR7 ATATATATAT GTATATATGT TACTCTACAT ATCAAATATC AAAACGTCTT TTATTTATTT AAAGTAAATT

Figure 4.14

band of the Canton S digest in Figure 4.12A) is replaced by a 6.7kb band, equivalent to a 2.4kb deletion.

To construct a genomic library, $\underline{w}^{IR8}$ DNA was cut with HindIII and ligated with HindIII-cut λNM1149. Resulting molecules were packaged *in vitro* and plated on NM514. Approximtely 50,000 plaques were screened with pI769 and four potential positive plaques were found. These were picked, plated out for single plaques and rescreened with pI769. Two positive clones were detected from which plate lysates and subsequently liquid lysates were made. The DNA was extracted and cut with HindIII to check the size of the inserts. One clone (λI426.1) had an insert of the correct size (6.7kb) but the other clone (λI426.2) had a much smaller insert (see Fig. 4.12C) and was not used for further experiments.

λI426.1 was subcloned into M13 by restricting the clone with SalI and ligating the fragments into SalI-cut mp18. White plaques recovered following transformation of NM522 were screened with a probe containing pCS155 and pCS156, to detect the small fragment expected to be present (fragment 'A' in Fig. 4.15B). Three positive plaques were found and these were picked, templates made and sequenced. All three of these M13 clones started from a SalI site within the lambda DNA. It was presumed that these fragments were due to a partial digestion and represented fragment 'B' shown in Figure 4.15B. To sequence the *white* DNA which was presumed to be at the far end of the fragments cloned in M13, the orientation of the fragments were reversed using the clone turn-around method described in Chapter 2.

Figure 4.15

Diagram to illustrate the expected structure of the $\underline{w}^{IR8}$ deletion and the actual structure.

A. The wild-type white sequence showing the three SalI sites, numbered 1, 2 and 3. The HindIII sites used to clone this region into λNM1149 are shown.

B. Expected structure of the lambda clone showing the proposed deletion between SalI sites 1 and 3. Open box represents white DNA and the single line represents the lambda arms. A SalI digest of the clone was expected to generate fragments 'A' and 'B'.

C. Actual structure of the $\underline{w}^{IR8}$ deletion, as deduced from clone λI426. Only SalI site number 3 remains in the white DNA, sites 1 and 2 are in the deleted region.

Symbols:- H, HindIII; S, SalI; B, BamHI.

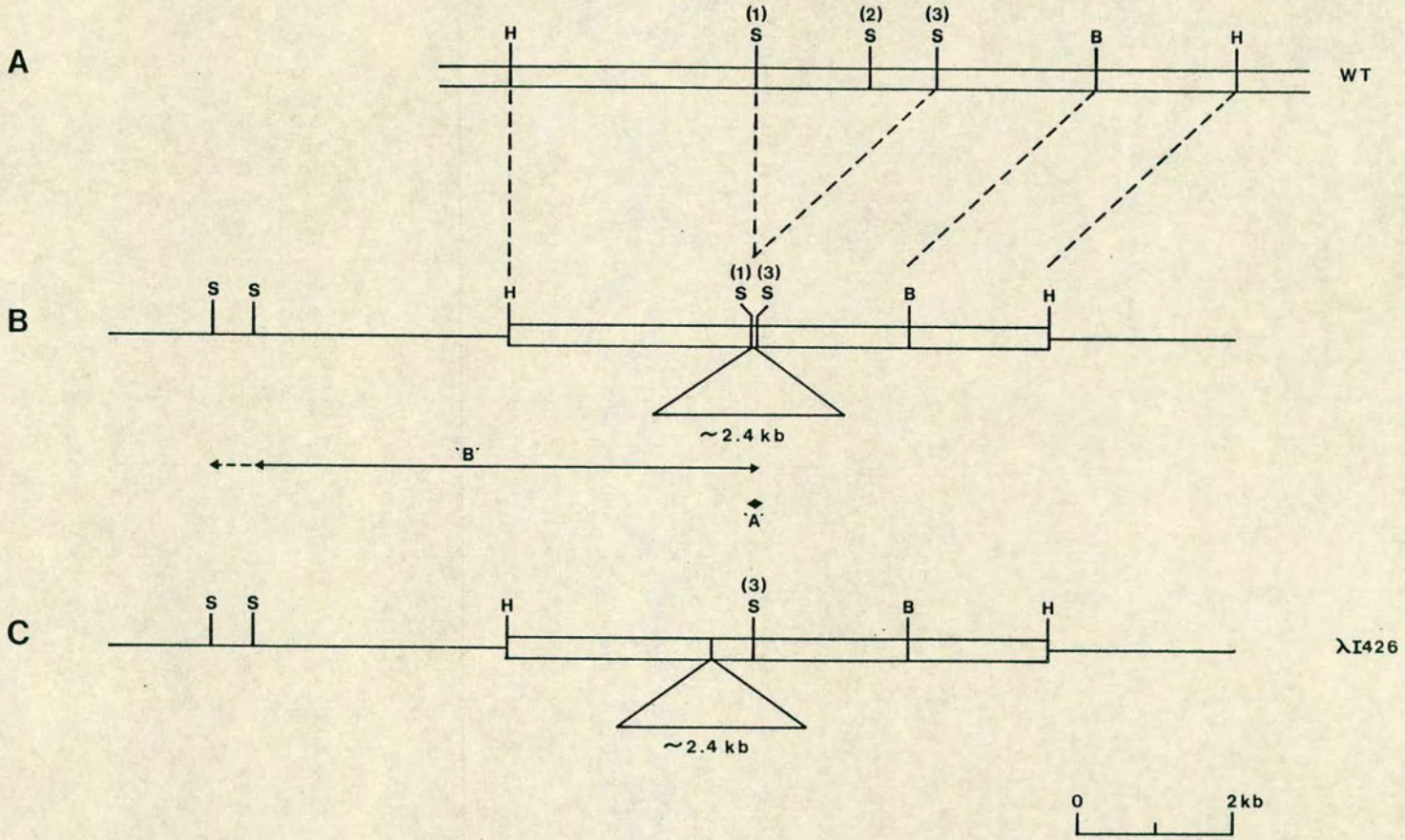


Figure 4.15

Twelve templates were prepared from each of the three reversed clones. Of these only one was found to start within recognisable white sequence. This clone started at the SalI site at the right of CS155 (site 3 in Fig. 4.15A). However, the sequence did not change from CS155 sequence to CS156 sequence after a few bases as expected but instead matched CS155 sequence for 516bp. At this point the sequence changed, but not to CS156 sequence. white-specific primers DF5 and DF6 were used to sequence this fragment (see Chapter 2). DF6 did not work however as this sequence occurs after the breakpoint.

The structure of this clone was interpreted in the following way. The deletion starts within CS155 and must end within I769 sequence at a distance from SalI site 1 (Fig. 4.15A) equivalent to the distance from SalI site 3 of the breakpoint. This would generate a fragment homologous to probe pI769 of the same size as the parental I769 fragment and hence would explain the result of the blot in Sang et al. (1984). The structure of the clone λI426.1 is therefore proposed to be as shown in Figure 4.15C.

To try and find the sequence on the other side of the breakpoint within I769 DNA the insert from pI769 was cloned into mp19 and sequenced from the SalI site. The site of the breakpoint should occur approximately 500bp from the SalI site and indeed was not found within the 350bp sequenced. A primer was made (DF7 - see Chapter 2) which matched the sequence beyond the breakpoint in w^{IR8} and hence should prime the I769 template. When this was tried however no priming

occurred. As the primer sequence did not match white DNA it was tested against lambda sequence using the UWGCG program FIND, and was found to correspond to bases 35074 to 35088 of lambda. Presumably this deletion occurred either in the initial subcloning from the lambda clone or in the clone-turn around.

In order to clone the genuine w^{IR8} deletion breakpoint λI426.1 was digested with SalI and HindIII and ligated into SalI/HindIII-cut mp19. Templates were made from white plaques found following NM522 transformation and these were sequenced from the SalI site and using primers DF5 and DF6. DF6 again did not prime. A breakpoint was found only 10bp from the first (between bases 12397 and 12398) and the sequence beyond it did not correspond to lambda. It was therefore presumed to be white sequence, although this has not been proven. A primer made specifically for I769 sequence, derived from the sequence already obtained, should enable the proposed site of the breakpoint within I769 to be sequenced.

An examination of the breakpoint, the sequence of which is shown in Figure 4.14, gives no indication of I factor activity. As with w^{IR7} however this cannot be excluded. The three deletion breakpoints, w^{IR7}, w^{IR8} and w^{IR8}/lambda, have all occurred within 30bp of one another. This suggests that this may be a deletion hotspot although the reason for this is not apparent. If this were true then it could be this feature of the sequence, rather than I factor activity, that was responsible for the generation of the w^{IR7} and w^{IR8} mutations.

4.10 Summary and Discussion

The sequence of both the ends and the flanking DNA has been determined for seven I factors and the following conclusions can be drawn. Firstly, there is a high degree of sequence conservation between different I factors. Within the first 110bp of the left end only one base was found to be variable (the third base). Where more sequence data was available only one other base was found to differ from the w^{IR1} sequence (base 125 of w^{IR5}). At the right end approximately 210bp of sequence was determined for the majority of the elements. No changes were found in any element. Because most of these sequences lie outside coding regions of the I factor it is perhaps surprising that the degree of conservation is so high. This may reflect a functional constraint on sequence changes suggesting that non-coding regions may be of some importance.

Secondly, the number of TAA triplets at the right hand end of the I factor is variable. So far elements with 4 (possibly), 5, 6 and 7 have been found. The nature of this tail will be discussed in the next chapter.

A third feature of note is the variation in size of the target site duplication between different elements. The majority of transposable elements in Drosophila duplicate a number of bases which is characteristic for each family. So far only the retroposons have been found to duplicate a variable number (Rogers, 1985; Weiner (et al., 1986). It is possible that retroposons do not possess a

specific integration mechanism but instead take advantage of random breaks in the chromosomes. It seems unlikely that I should integrate *via* such a mechanism however, as an insertion hotspot is known. More likely, the mechanism whereby staggered nicks are generated may not be as precise as for other elements.

The presence of an insertion hotspot implies that some feature of the sequence is recognised by the integration mechanism. The sequences of the duplications flanking the seven elements (Fig. 4.3) were compared in order to detect any homology between them and to try to derive a consensus sequence. With the exception of $\underline{w}^{IR1}$, $\underline{w}^{IR3}$ and $\underline{w}^{IR4}$ however they are quite different from one another and the only conclusion which can be made is that all the sequences tend to be A/T rich. It is possible that it is not the sequence of the target site itself that is recognised for integration but a sequence further away or possibly secondary structure of the DNA.

Finally, the positions of the I factor insertions and the deletion breakpoints have been mapped precisely within the *white* gene and some correlation of the genotype and phenotype of these mutations can be made. The white-eyed phenotype of $\underline{w}^{IR7}$ and $\underline{w}^{IR8}$ is readily explainable by the deletion of *white* coding sequences. All the insertions have occurred within non-coding regions and hence their effects are probably due to interference with the correct processing of the *white* RNA. Some eye colour remains in these flies which could be explained either by a low level of RNA being correctly processed or all the RNA being aberrantly processed and coding for a product which

has only partial activity. The effects seem to be different according to the position of the I factor within the gene. Strains $\underline{w}^{IR1}$, $\underline{w}^{IR3}$ and $\underline{w}^{IR4}$ have brown eyes but the degree of pigmentation is temperature sensitive. The eyes are lighter when the flies are placed at 25°C than they are at 20°C. This could be explained by a product being made from the white gene which has a temperature sensitive conformation, possibly due to the incorporation into the protein of some I factor-encoded amino acids. Alternatively the RNA itself may be sensitive to temperature.

The other strains ($\underline{w}^{IR2}$, $\underline{w}^{IR5}$ and $\underline{w}^{IR6}$) also have brown eyes, but in these flies the eye colour is not temperature sensitive. The effect of the $\underline{w}^{IR2}$ insertion is of particular interest as it causes partial restoration of eye colour in $\underline{w}^{1}$ flies. The $\underline{w}^{1}$ insertion lies between the white promoter and the coding sequences. It seems unlikely that the I factor itself could be providing an alternative promoter as there would still be several kb of foreign DNA (the $\underline{w}^{1}$ element) between the promoter and the start of the white coding region. It is possible that a low level of white transcripts are spliced to remove the insertions, but the mechanism whereby a secondary insertion could cause such an event is unknown.

CHAPTER 5

Discussion

5.1 Summary

The DNA sequence of one complete and functional I factor has been determined. In addition the ends of several other I factors have been cloned and sequenced.

The I factor is 5.4kb long, has no terminal repeats and is flanked by a target site duplication of variable size. At the right hand end (the 3' end) of the element there is a run of TAA triplets, the number varying between elements. Two long open reading frames are found within the sequence, both on the top strand. The shorter of these (ORF1) encodes a protein which contains homology to a sequence motif characteristic of retroviral *gag* proteins and which may be involved in binding nucleic acids. The larger open reading frame (ORF2) encodes a protein with features common to reverse transcriptases.

In structure the I factor resembles the class of transposable elements known as retroposons. These elements are thought to transpose *via* reverse transcription of an RNA intermediate. The finding of a reverse transcriptase-like protein encoded by the I factor provides further evidence for the relationship between the I factor and retroposons. Particularly good homology has been found between the putative reverse transcriptases of the I factor and mammalian LINE elements, leading to the suggestion that the I factor may represent a *Drosophila* equivalent of a LINE.

5.2 The TAA tail

Retroposons are characterised by the possession of an A-rich tail. The TAA triplets found at the right-hand end of the I factor seem unlikely to constitute a poly A tail of this sort because of the repetitive nature of the TAA motif. It is hard to see how this could be a degenerative poly A tail, especially as all the I factors looked at here had tails of the same structure differing only in the number of triplets. In addition there is no polyadenylation site at an appropriate distance from the right-hand end of the I factor which could be used to poly-adenylate a full length message. The closest polyadenylation site is 79bp upstream of the first TAA triplet and hence is too far away. It is possible that the TAA repeats are an integral part of the I factor and that the number increases or decreases as a result of unequal recombination between elements. Alternatively the number of triplets could vary as a result of integration of the element into genomic DNA. Integration probably occurs at staggered nicks and during the subsequent polymerisation and ligation to generate an integrated element flanked by direct repeats slippage could occur and generate or remove TAA triplets. Another possibility is that the triplets are not an integral part of the element but are added to transposition intermediates in a manner analogous to that of telomere expansion in Tetrahymena (Greider & Blackburn, 1985). This would require an enzyme system not yet identified in Drosophila.

5.3 The mechanism of transposition

Based on observations from the sequence data a possible transposition mechanism can be proposed. This mechanism is, as yet, highly

speculative.

The first step of a transposition cycle involving reverse transcriptase would be the generation of a full length RNA. It is assumed that the strand copied would be the same as that used for expression of the two open reading frames as this is in keeping with the proposed mechanism for retroposons as well as for retroviruses and copia-like elements. Transcription would therefore start at the left-hand end of the element and terminate at the right-hand end. It is unclear which of the three RNA polymerases (I, II or III) would be used to transcribe the I factor. As has been discussed elsewhere an internal promoter would be essential for an element to be capable of more than one transposition event. Polymerase III transcription units have to date all been found to contain internal promoters, known as the "A" and "B" box sequences (Galli et al., 1981; Rogers, 1985) within the first 200 bp of the gene. These boxes are not present within the first 200bp of the I factor so it would appear that polIII is not used. In addition several runs of four or more T residues are found in the non-coding strand, known to be polIII termination signals (Bogenhagen & Brown, 1981). The promoters for both polI and polII transcription units both precede the gene and hence are not included in the transcript. It would seem that for an element with no internal promoter to be capable of transposition it would have to insert fortuitously after the promoter of a gene. This is not the case for w^{IR1} however as the element has not inserted near the white promoter, and is in the wrong orientation for a white transcript to read through the I factor on the correct strand. It is possible that

the I factor carries a novel internal polI or polII promoter, or a polIII promoter which does not conform to the canonical "A" and "B" boxes. In any of these cases the promoter would be carried with the element when it transposed.

Loeb et al. (1986) have sequenced one complete LINE from the mouse (L1Md.A2) and the 5' (left-hand) end of a second element. These LINEs contain $4^2/3$ and $1^2/3$ copies of a 208bp direct tandem repeat respectively at their 5' ends. They propose that these repeats contain promoter sequences and that transcription starts from within one of the repeats. Although the L1Md sequence preceding the promoter will be lost upon transcription, Loeb et al. (1986) suggest that recombination between elements containing different numbers of repeats could restore extra copies. Human and rat LINEs and the I factor contain no such repeats and hence promotion must be explained by an alternative means such as a novel internal promoter.

Once the full length RNA has been made, the second step of the transposition cycle would be reverse transcription of the RNA. This process requires a primer, which could be provided by the 3' end of the RNA folding back on itself or by the binding of another RNA molecule, such as a tRNA, to the 3' end. Both these possibilities have the disadvantage that sequence would be lost from the 3' end. It is possible that transcription does not stop at the end of the I factor but carries on into the adjacent sequence until an appropriate stop signal is reached. If this is in an A/T rich region then the end of the RNA may be able to pair with the TAA triplets at the end of the I

factor and thus act as a primer. This sort of mechanism has been proposed for the Alu-type retroposons (Rogers, 1985). It has already been noted that I factors seem to insert preferentially into A/T rich regions, this may be as a result of requirements for transposition.

The next step would either be integration of the DNA/RNA hybrid into the genome, or removal of the RNA and synthesis of the second DNA strand prior to integration. For the latter mechanism to be employed a second primer would be needed. Again this would have the disadvantage that sequence would be lost, this time from the left end of the element. It should be noted that LINE elements with truncations at the left hand end are frequently found (Singer & Skowronski, 1985) and defective I elements have more often lost sequence from the left end than the right end (Bucheton & Vaury, unpublished observations). Whether this is as a consequence of transcription starting in the wrong place or reverse transcription terminating prematurely is not known, but it would argue against the use of a primer to generate a second DNA strand as frequently the primer binding site would be missing. The former mechanism (integration of a DNA/RNA hybrid) would seem to be the more viable alternative, therefore.

The final step of the transposition cycle, then, would be integration of the element into the genome. This may be an element encoded function analogous to the retroviral integrase enzyme, although as yet no such gene has been identified in the I factor. Integration is unlikely to occur by a random mechanism, such as at random nicks in the chromosome, as an insertion hotspot is known.

Clearly much experimental work is necessary to elucidate the mechanism of transposition. In particular it would be advantageous to isolate the full length RNA to see whether the 5' and 3' ends correspond to the ends of the I factor or whether transcription includes flanking sequences. The major problem may be to isolate sufficient quantities of RNA. It is likely that full length transcripts are only made in the germ cells of dysgenic females, as it is known that the elements can only transpose under such conditions.

Each step of the transposition cycle of copia-like elements has been identified and all transposition intermediates have been detected (Archipova et al., 1986). The mechanism in principle is analagous to that of retroviruses. The mechanism employed by the I factor must be quite different as I has no terminal repeats, features crucial to the generation of first and second strand DNA molecules of retroviruses.

5.4 The role of an I factor repressor

It has been established that the transposition of I factors within I strains, if it occurs at all, occurs at a very low frequency. In contrast, when I factors are introduced into R strains transposition occurs at a relatively high frequency. Two hypotheses have been advanced to explain this, the active induction of transposition by factors present only in R strains or active repression by factors present only in I strains. The finding of a possible nucleic acid binding protein encoded by the I factor seems to favour the latter hypothesis. There are several possibilities for how such a repressor

could work. If, for example, it could bind to DNA, transposition could be prevented by binding of the repressor to the reverse transcriptase promoter or to the promoter for the full length RNA. It is possible however that preventing production of reverse transcriptase alone would not prevent transposition completely as other transposable elements (the copia-like) produce such an enzyme which may reverse transcribe other RNAs. If the repressor binds RNA then binding to the full length RNA to prevent reverse transcription would inhibit transposition. The role of the putative repressor could be investigated using a transformation system. Preliminary data obtained from injection of functional I factors (cloned into the P element vector) into R strain embryos suggests that transformed flies exhibit some properties of I strains, such as induction of dysgenesis when crossed with an R strain (Pritchard, Dura, Pelisson & Finnegan, unpublished data). The establishment of I cytotype in the transformants has not yet been demonstrated, however, but would be expected as the introduction of I factors into R cytoplasm (a dysgenic cross) leads to rapid cytotype switch to the I type. The potential repressor gene could be mutated, for example, by altering the conserved $CX_2CX_4HX_4C$ motif, and the element then injected into R strain embryos. Subsequent transposition of the I factor and no establishment of the I cytotype (assuming that this can be demonstrated with an unmutated I factor) would be some evidence for the involvement of the gene product in regulation of I factor activity.

The alternative hypothesis, that R strains induce transposition,

cannot be ruled out however. It is possible that the repressor does not repress the I factor itself but represses the inducer. This may be an unnecessarily complicated explanation, especially in view of the proposals that R strains are derived from I strains by loss of functional I factors, or that I strains are derived from R strains by restoration of I factor function.

5.5 Germ line and female specificity of I factor transposition

One feature in common between the IR and PM systems of hybrid dysgenesis is the restriction of transposition to the cells of the germ line. In IR dysgenesis transposition occurs only in the germ line of hybrid females.

In the PM system Laski et al. (1986) performed experiments to distinguish between the possibilities that tissue-specific promotion or splicing were responsible. They found that the first three open reading frames of the P factor (ORF0, ORF1 and ORF2) are spliced together in somatic tissue but that the splice between these three open reading frames and the fourth (ORF3) only occurs in the germ line. As all four open reading frames are required for a functional transposase, such an enzyme will be found only in the germ line.

The same possibilities can be applied to the I factor. The two open reading frames could be spliced together, either from an RNA made from a promoter immediately preceding ORF1 and ending just after ORF2, or alternatively they could be spliced from the full length RNA. This seems unlikely as a computer search found no suitable splice donor and

acceptor sites (Mount, 1982) within the open reading frames. Splicing cannot be ruled out however as it has not been proved that variations of the splicing signals cannot sometimes be recognised.

One of the putative I factor promoters, either for ORF2 or for the full length RNA, may be tissue specific. If this is the case then I factor transcripts would not be expected to be found in somatic tissue. This is difficult to investigate however as there are so many defective I elements in all D. melanogaster strains which could give rise to non-functional RNA, either from promoters within the remaining sequences or from promoters in flanking sequences. Unfortunately there are no R strains equivalent to the M strains which completely lack P-homologous sequences. It may be possible to inject a functional I factor into one of the more distantly related Drosophila species which contain no I factor homology (Bucheton et al., 1986) and to investigate the distribution of I factor transcripts in the tissues of transformants.

Another possibility for tissue specificity is the presence of some molecule in the germ cells only, such as an RNA which could act as a primer for reverse transcription.

Transposition occurs only in the germ line of hybrid females. Lavige (1986) has suggested that events which occur during oogenesis in SF females are responsible for the developmental arrest of their progeny. It is probable that some aspect of the biochemistry of developing oocytes facilitates I factor transposition.

5.6 Hybrid dysgenesis in other species

The apparent relationship between the I factor and mammalian LINE elements raises the interesting question of whether hybrid dysgenesis can and does occur in other species. In theory this would require two kinds of strain, analagous to the I and R strains of *Drosophila melanogaster*, i.e. one strain possessing autonomous LINE elements and another strain possessing only defective LINEs. Is there any evidence for such strains existing among mammals? Every mammalian species studied to date has one major family of LINEs, present in many thousands of copies per genome. As yet only one new insertion of a LINE element has been witnessed, in the dog (Katzir *et al.*, 1985), but some, if not all, of the elements are thought to have transposed to their present location because they are flanked by direct repeats. Recently Skowronski and Singer (1985) have detected a 6.5kb RNA in human tissue culture cells which is homologous to LINE elements. This RNA has the structure expected of a transposition intermediate but it is not known if it is being made from a functional LINE. So, there is little evidence as yet to suggest that LINEs can still transpose, except at a very low frequency as only one transposition event has been witnessed. This could be because all members of a species contain functional LINEs which, like I factors in an inducer strain, are repressed and transpose rarely. Alternatively, all members of a species may contain only defective LINEs, i.e. they are all the equivalent of an R strain.

There is one example, however, in mice, in which a strain exhibits

some characteristics of hybrid dysgenesis. It has been shown that female mice of the DDK strain show lower fertility when mated with males of another strain than when mated with DDK males. In contrast, females of other strains show normal fertility when crossed with DDK males (Wakasugi, 1973). Reduction in fertility was found to be due to embryo death at a very early stage of development. Wakasugi (1973) proposed that this was due to some substance present in the cytoplasm of DDK females which somehow interacted with the chromosomes of males of different strains, but not DDK males. It was later demonstrated (Wakasugi, 1974) that the chromosomal factor carried by the spermatazoa, and the cytoplasmic factor carried by the egg, were both determined by autosomal genes and furthermore that these genes were at the same locus or very closely linked. Although other suggestions have been put forward to explain this interaction (Wakasugi, 1974) it could be explained as a sort of hybrid dysgenesis. DDK mice could be analagous to a D. melanogaster R strain, and the strains which interact with DDK mice could be the equivalent of I strains. The induction of sterility is compatible with a dysgenesis system, although no other features, such as a higher mutation frequency, have been reported. Mutations may not be seen, however, as the frequency of I factor-induced mutations is only approximately 100-fold higher than the spontaneous mutation rate.

It should be noted that there is an important difference between Drosophila hybrid dysgenesis and the potential system in the mouse. In Drosophila transposition is limited to the germ line, hence it is the hybrid progeny which exhibit reduced fertility. In the mouse

reduced fertility is seen in the first cross, hence if transposable elements are involved they must be able to transpose in somatic tissues.

Experiments are underway to see whether DDK mice contain LINE elements (D. Finnegan, unpublished data), and if they do whether there are any obvious differences between DDK mice and other strains, such as a disparity in element copy number, or a lack of full length elements.

5.7 The variety of reverse transcriptases

Proteins showing homology to reverse transcriptase have been identified from a variety of different sources. The enzyme was first discovered in retroviruses and it was proposed that reverse transcriptase was originally a cellular gene which had been incorporated into a viral genome when retroviruses were first being formed (Temin, 1980; Flavell, 1981). Since this first discovery genes coding for reverse transcriptases have been identified in *Drosophila copia*-like transposable elements, yeast Ty transposable element, mammalian LINEs, fungal class II mitochondrial introns, cauliflower mosaic virus, hepatitis B virus and now in the *Drosophila* I factor. One question raised by these findings is, are all these reverse transcriptases related and undergoing divergent evolution, or did the genes arise independently and are undergoing convergent or parallel evolution because they carry out a similar function? It is clear that at least in the case of retroviruses, *copia*-like elements and Ty elements, the similarities go too far beyond the mere possession of a

related gene for there to be no closer relationship. The status of the other reverse transcriptase genes is less clear, but the degree of homology between them is more suggestive of a distant common ancestor than of independent origins. The discovery of a cellular reverse transcriptase, the putative common ancestor, has not yet been made but if such a gene were to be found it would make the possibility of a common origin seem more likely.

Another point of debate is whether reverse transcriptase has any function within the cell other than for viral replication or element transposition. It is difficult to think of any process within the cell for which reverse transcriptase would be required. It is possible that reverse transcriptase evolved as part of a transposable element and has no direct role within the cell, and probably never had. The consequences for a cell possessing a reverse transcriptase may be trivial or possibly even advantageous. Baltimore (1985) has suggested that such an enzyme activity could play a part in genome evolution. Assuming that reverse transcriptase could occasionally copy cellular RNAs, the resulting cDNAs could, by recombination and gene conversion, result in genes which contained fewer, or no, introns. These genes could be functional as the promoter would still exist. This is distinct from integration of the cDNA itself into the genome to generate a pseudogene which would not possess a promoter (except on rare occasions).

The role of reverse transcriptase within fungal mitochondria is unclear. The open reading frames containing the reverse transcriptase

homology are clearly under selective pressure as usually intron sequences evolve at a relatively high rate. There may be reverse transcriptase activity in the mitochondria. Several instances of clean excision of introns have been noted in S. cerevisiae, presumed to be due to recombination and gene conversion between the mitochondrial gene and the cDNA (Flavell, 1985). In addition, circular DNA molecules have been isolated from P. anserina which are identical to one of the mitochondrial introns which contains reverse transcriptase homology (Michel & Lang, 1985). As yet no biological function has been ascribed to this enzyme, if indeed it is active within mitochondria. It is not known whether this gene is cellular (the putative "ancestral" gene) or whether it was carried into the mitochondria by a viral horizontal transmission event.

Toh et al. (1985) have suggested a common ancestor not only for retroviruses and copia-like elements but also for the hepatitis B group viruses and cauliflower mosaic viruses, based on protein sequence homologies. They propose the hepatitis B group viruses to be the most divergent as they show homology only in the reverse transcriptase region. CaMV on the other hand shows a closer relationship having homology to retroviral reverse transcriptase, protease and nucleic acid binding domain (the $CX_2CX_4HX_4C$ motif) (Toh et al., 1985; Covey, 1986). The double stranded, circular DNA form of retroviruses (the integration substrate) could be analagous to these double stranded DNA viruses except that both HBV and CaMV lack LTR structures.

If the majority of reverse transcriptase-containing sequences can be explained by a common ancestor, could the I factor and mammalian LINEs be fitted into such a model? Homology has been found within the I factor to both the retroviral reverse transcriptase and also to the gag-specific nucleic acid binding motif. It is therefore tempting to suggest that once again there is a common ancestor and that the two types of element evolved in different ways, utilizing the same enzyme. Alternatively one type of element could have evolved from the other, either by loss of LTRs to form the I factor, or more probably, by evolution of LTRs to form retroviruses. LINEs show less homology to retroviruses than does the I factor, for example there is no nucleic acid binding domain homology. Nevertheless a similar event could have occurred, or the event could have occurred only once to generate a LINE element which was the ancestor of both the mammalian LINEs and the Drosophila I factor.

Although there appears to be an evolutionary relationship between these viruses and transposable elements this could also be explained by several independent events whereby a reverse transcriptase gene was combined with other genes (nucleic acid binding proteins, proteases, etc.) to form various different viruses or transposable elements. Successful combinations may then have proliferated, and could have arisen more than once in different forms.

5.8 Concluding Remarks

The sequencing of a functional I factor has generated a good deal of information, including the structure of the element, coding capacity

and probable nature of the gene products. This has allowed the proposal of a likely transposition mechanism. There is still much that needs to be done to check the validity of these proposals however. I factor-specific transcripts need to be isolated and characterised, for example by primer extension analysis to map the start of transcription of the full length RNA and also of the two open reading frames. Likewise the products of the two open reading frames need to be isolated. The product of ORF2 needs to be tested for reverse transcriptase activity. The product of ORF1 needs to be tested for nucleic acid binding activity and, if it possesses this, potential binding sites on the DNA or RNA should be identified, by footprint analysis or S_1 protection.

An embryo transformation system for the I factor would be invaluable for studying the effects of _in vitro_ manufactured I factor mutants. Promoter mutations, open reading frame mutations, mutations of the TAA triplets (such as complete removal) could be studied, as well as the nature of tissue specificity. If the tissue specificity could be overcome it may be possible to induce transposition in _Drosophila_ tissue culture cells, from which greater amounts of transcripts, proteins and possibly transposition intermediates could be isolated than from the flies themselves.

Some of these experiments are already underway in this laboratory and should result in a much fuller understanding of this unusual transposable element.

REFERENCES

ARBER, W., ENQUIST, L., HOHN, B., MURRAY, N.E. and MURRAY, K. (1983). "Experimental Methods for Use with Lambda" in "Lambda II" eds. Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A. pp.433-466.

ARKHIPOVA, I.R., GORELOVA, T.V., ILYIN, Y.V. and SCHUPPE, N.G. (1984). Nuc.Acids Res. 12, 7533-7548.

ARKHIPOVA, I.R., MAZO, A.M., CHERKASOVA, V.A., GORELOVA, T.V., SCHUPPE, N.G. and ILYIN, Y.V. (1986). Cell 44, 555-563.

BELLET, A.J.D., BUSSE, H.G. and BALDWIN, R.L. (1971). "Tandem Genetic Duplications of Phage Lambda" in "Bacteriophage Lambda", ed. A.D. Hershey, pp.501-513.

BENTON, W.D. and DAVIS, R.W. (1977). Science 196, 180-182.

BIGGIN, M.D., GIBSON, T.J. and HONG, G.F. (1983). Proc.Nat.Acad.Sci. USA 80, 3963-3965.

BINGHAM, P.M., KIDWELL, M.G. and RUBIN, R.M. (1982). Cell 29, 995-1004.

BORCK, K., BEGGS, J.D., BRAMMAR, W.J., HOPKINS, A.S. and MURRAY, N.E. (1976). Mol.Gen.Genet. 146, 199-207.

BREGLIANO, J.-C. and KIDWELL, M.G. (1983). "Hybrid Dysgenesis Determinants" in "Mobile Genetic Elements", ed. J. Shapiro, pp.363-410.

BROOKFIELD, J.F.Y., MONTGOMERY, E. and LANGLEY, C.H. (1984). Nature 310, 330-332.

BUCHETON, A. (1978). Heredity 41, 357-369.

BUCHETON, A. (1979a). Genetics 93, 131-142.

BUCHETON, A. (1979b). Biol.Cell. 34, 43-49.

BUCHETON, A. and PICARD, G. (1978). Heredity 40, 207-223.

BUCHETON, A., LAVIGE, J.M., PICARD, G. and L'HERETIER, P.H. (1976). Heredity 36, 305-314.

BUCHETON, A., PARO, R., SANG, H.M., PELISSON, A. and FINNEGAN, D.J. (1984). Cell 38, 153-163.

BUCHETON, A., SIMONELIG, M., VAURY, C. and CROZATIER, M. (1986). Nature 322, 650-652.

BURTON, F.H., LOEB, D.D., VOLIVA, C.F., MARTIN, S.L., EDGELL, M.H. and HUTCHISON, C.A. (1986). J.Mol.Biol. 187, 291-304.

COLLINS, J.F. and COULSON, A.F.W. (1986). "Molecular Sequence Comparison

and Alignment" in "Nucleic Acid and Protein Sequence Analysis: a Practical Approach" eds. Bishop, M. and Rawlings, C., in press.

COPELAND, T.D., OROZLAN, S., KALYANARAMAM, V.S., SARNGADHARAN, M.G. and GALLO, R.C. (1983). FEBS Lett. 162, 390-395.

COVEY, S.N. (1986). Nuc.Acids Res. 14, 623-633.

D'AMBROSIO, E., WAITZKIN, S.D., WITNEY, F.R., SALEMME, A. and FURANO, A.V. (1986). Mol.Cell.Biol. 6, 411-424.

DANIELS, S.B., STRAUSBAUGH, L.D., EHRMAN, L. and ARMSTRONG, R. (1984). Proc.Nat.Acad.Sci. USA 81, 6794-6797.

DAWID, I.B., OLONG, E.O., DI NOCERA, P.P. and PARDUE, M.L. (1981). Cell 25, 399-408.

DEININGER, P.I. (1983). Anal.Biochem. 129, 215-223.

DEVEREUX, J., HAEBERLI, P. and SMITHIES, O. (1984). Nuc.Acids Res. 12, 387-395.

DICKSON, C., EISENMAN, R. and FAN, H. (1985). "Protein Biosynthesis and Assembly" in "RNA Tumour Viruses", eds. Weiss, R., Teich, N., Varmus, H. and Coffin, J., pp.135-145.

DI NOCERA, P.P. and DAWID, I.B. (1983). Nuc.Acids Res. 11, 5475-5482.

DI NOCERA, P.P., GRAZIANI, F. and LAVORGNA, G. (1986). Nuc.Acids Res. 14, 675-691.

EFSTRATIADIS, A., POSAKONY, J.W., MANIATIS, T., LAWN, R.M., O'CONNELL, C., SPRITZ, R.A., DERIEL, J.K., FORGET, B.G., WEISSMAN, S.M., SLIGHTOM, J.L., BLECHL, A.E., SMITHIES, O., BARALLE, F.E., SHOULDERS, C.C. and PROUDFOOT, N.J. (1980). Cell 21, 653-668.

EMORI, Y., SHIBA, T., KANAYA, S., INOUYE, S., YUKI, S. and SAIGO, K. (1985). Nature 315, 773-776.

ENGELS, W.R. (1979a). Genet.Res.Camb. 33, 219-236.

ENGELS, W.R. (1979b). Proc.Nat.Acad.Sci. USA 76, 4011-4015.

ENGELS, W.R. (1981). Cold Spring Harb.Symp.Quant.Biol. 45, 561-565.

ENGELS, W.R. (1983). Ann.Rev.Genet. 17, 315-344.

ENGELS, W.R. and PRESTON, C.R. (1979). Genetics 92, 161-174.

ENGELS, W.R. and PRESTON, C.R. (1984). Genetics 107, 657-678.

FAWCETT, D.H., LISTER, C.K., KELLETT, E. and FINNEGAN, D.J. (1986). Cell 47, 1007-1015.

FINNEGAN, D.J. and FAWCETT, D.H. (1986). "Transposable Elements in Drosophila melanogaster" in "Oxford Surveys on Eukaryotic Genes" 3, 1-62.

FINNEGAN, D.J., RUBIN, G.M., YOUNG, M.W. and HOGNESS, D.S. (1978). Cold Spring Harb.Symp.Quant.Biol. 42, 1053-1063.

FLAVELL, A.J. (1984). Nature 310, 514-516.

FLAVELL, A.J., RUBY, S.W., TOOLE, J.J., ROBERTS, B.E. and RUBIN, G.M. (1980). Proc.Nat.Acad.Sci. USA 77, 7107-7111.

FLAVELL, A.J., LEVIS, R., SIMON, M.A. and RUBIN, G.M. (1981). Nuc.Acids Res. 9, 6279-6291

FRISCHAUF, A.-M., LEHRACH, H., POUSTKA, A. and MURRAY, N. (1983). J.Mol.Biol. 170, 827-842.

GOUGH, J.A. and MURRAY, N.E. (1983). J.Mol.Biol. 166, 1-19.

GREEN, M.M. (1976) in "The Genetics and Biology of Drosophila" eds. Ashburner, M. and Novitsky, E. Vol. 16, pp.929-946.

HANAHAN, D. (1983). J.Mol.Biol. 166, 557-580.

HATTORI, M., KUHARA, S., TAKENAKA, O. and SAKAKI, Y. (1986). Nature 321, 625-628.

HAYNES, S.R. and JELINEK, W.R. (1981). Proc.Nat.Acad.Sci. USA 78, 6130-6134.

HOUCK, C.M., RINEHART, F.P. and SCHMID, C.W. (1979). J.Mol.Biol. 132, 289-306.

HU, N. and MESSING, J. (1982). Gene 17, 271-277.

ISING, G. and BLOCK, K. (1981). Cold Spring Harb.Symp.Quant.Biol. 45, 527-549.

ISING, G. and BLOCK, K. (1984). Mol.Gen.Genet. 196, 6-16.

JELINEK, W.R., TOOMEY, T.P., LEINWAND, L., DUNCAN, C.H., BIRO, P.A., CHOUDARY, P.V., WEISSMAN, S.M., RUBIN, G.M., HOUCK, C.M., DEININGER, P.L. and SCHMID, C.W. (1980). Proc.Nat.Acad.Sci. USA 77, 1398-1402.

KARESS, R.E. and RUBIN, G.M. (1984). Cell 38, 135-146.

KIDWELL, M.G. (1979). Genet.Res.Camb. 33, 205-217.

KIDWELL, M.G. (1982). "Intraspecific Hybrid Sterility" in "The Genetics and Biology of Drosophila" eds. Ashburner, M., Carson, H. and Thompson, J.N.Jr., Vol.3c.

KIDWELL, M.G. (1983). Proc.Nat.Acad.Sci. USA 80, 1655-1659.

KIDWELL, M.G. and KIDWELL, J.F. (1976). Genetics 84, 333-351.

KIDWELL, M.G. and NOVY, J.B. (1979). Genetics 92, 1127-1140.

KIDWELL, M.G., KIDWELL, J.F. and SVED, J.A. (1977). Genetics 86, 813-833.

KLEIN, B. and MURRAY, K. (1979). J.Mol.Biol. 133, 289-294.

KOZAK, M. (1986). Cell 44, 283-292.

KUGIMIYA, W., IKENAGA, H. and SAIGO, K. (1983). Proc.Nat.Acad.Sci. USA 80, 3193-3197.

LANSMAN, R.A., STACEY, S.N., GRIGLIATTI, T.A. and BROCK, H.W. (1985). Nature 318, 561-563.

LASKEY, R.A. and MILLS, A.D. (1977). FEBS Lett. 82, 314.

LASKI, F.A., RIO, D.C. and RUBIN, G.M. (1986). Cell 44, 7-19.

LAVIGE, J.M. (1986). Biol.Cell, in press.

LAVIGE, J.M. and LECHER, P. (1982). Biol.Cell. 44, 9-14.

LINDSLEY, D.L. and GRELL, E.H. (1968). "Genetic Variation of Drosophila melanogaster". Carnegie Inst. Washington, D.C.

LOEB, D.D., PADGETT, R.W., HARDIES, S.C., SHEHEE, W.R., COMER, M.B., EDGELL, M.H. and HUTCHISON, C.A. (1986). Mol.Cell.Biol. 6,

168-182.

MESSING, J. (1983). Meth.Enz. 101, 20-78.

MESSING, J. and VIEIRA, J. (1982). Gene 19, 269-276.

MICHEL, F. and LANG, B.F. (1985). Nature 316, 641-643.

MIZUSAWA, S. and WARD, D.F. (1982). Gene 20, 317-322.

MOUNT, S.M. and RUBIN, G.M. (1985). Mol.Cell.Biol. 5, 1630-1638.

MURRAY, N.E. (1983). "Phage Lambda and Molecular Cloning" in "Lambda II", eds. Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A. pp.395-432.

MURRAY, N.E., BRAMMAR, W.J. and MURRAY, K. (1977). Mol.Gen.Genet. 150, 53-61.

NORRANDER, J., KEMPE, T. and MESSING, J. (1983). Gene 26, 101-106.

O'HARE, K. (1985). Trends Genet. 1, 250-254.

O'HARE, K. and RUBIN, G.M. (1983). Cell 34, 25-35.

O'HARE, K., MURPHY, C., LEVIS, R. and RUBIN, G.M. (1984). J.Mol.Biol. 180, 437-455.

PARO, R., GOLDBERG, M.L. and GEHRING, W.J. (1983). EMBO J. 2, 853-860.

PEIFER, M. and BENDER, W. (1986). EMBO J., in press.

PELISSON, A. (1981). Mol.Gen.Genet. 183, 123-129.

PICARD, G. (1976). Genetics 83, 107-123.

PICARD, G. (1978). Biol.Cell. 31, 245-254.

PICARD, G. and L'HERETIER, P. (1971). Dros.Inf.Serv. 46, 54.

PICARD, G., BUCHETON, A., LAVIGE, J.M. and PELISSON, A. (1976). C.R. Hebd.Seances Acad.Sci.Ser. D 282, 1813-1816.

PICARD, G., LAVIGE, J.M., BUCHETON, A. and BREGLIANO, J.C. (1977). Biol.Cell. 29, 89-98.

PICARD, G., BREGLIANO, J.C., BUCHETON, A., LAVIGE, J.M., PELISSON, A. and KIDWELL, M.G. (1978). Genet.Res.Camb. 32, 275-287.

POTTER, S.S. (1982). Nature 297, 201-204.

POTTER, S.S., TRUETT, M., PHILLIPS, M. and MAHER, A. (1980). Cell 20, 639-647.

PROUST, J. and PRUDHOMMEAU, C. (1982). Mutat.Res. 95, 225-235.

RIO, D.C., LASKI, F.A. and RUBIN, G.M. (1986). Cell 44, 21-32.

ROGERS, J.E. (1983). Nature 301, 460.

ROGERS, J.E. (1985). Int.Rev.Cytol. 93, 188-280.

RONSSERAY, S., ANXOLABEHERE, D. and PERIQUET, G. (1984). Mol.Gen. Genet. 196, 17-23.

RUBIN, G.M. and SPRADLING, A.C. (1982). Science 218, 348-353.

RUBIN, G.M., KIDWELL, M.G. and BINGHAM, P.M. (1982). Cell 29, 987-994.

SAIGO, K., KUGIMIYA, W., MATSUO, Y., INOUYE, S., YOSHIOKA, K. and YUKI, S. (1984). Nature 312, 659-661.

SANG, H.M., PELISSON, A., BUCHETON, A. and FINNEGAN, D.J. (1984). EMBO J. 3, 3079-3085.

SANGER, F., COULSON, A.R., SMITH, A.J.H. and ROE, B.A. (1980). J.Mol.Biol. 143, 161-178.

SCHERER, G., TELFORD, J., BALDARI, C. and PIRROTTA, V. (1981).

Dev.Biol. 86, 438-447.

SCHERER, G., TSCHUDI, C., PERERA, J., DELIUS, H. and PIRROTTA, V. (1982). J.Mol.Biol. 157, 435-452.

SHIBA, T. and SAIGO, K. (1983). Nature 302, 119-124.

SIMMONS, M.J. and BUCHOLZ, M.B. (1985). Proc.Nat.Acad.Sci. USA 82, 8119-8123.

SIMMONS, M.J. and KARESS, R.E. (1985). Dros.Inf.Serv. 61, 2-7.

SINGER, M.F. (1982). Cell 28, 433-434.

SINGER, M.F. and SKOWRONSKI, J. (1985). TIBS 10, 119-122.

SKOWRONSKI, J. and SINGER, M.F. (1985). Proc.Nat.Acad.Sci. USA 82, 6050-6054.

SMITH, G.E. and SUMMERS, M.D. (1980). Anal.Biochem. 109, 123-129.

SPRADLING, A.C. and RUBIN, G.M. (1982). Science 218, 341-347.

STADEN, R. (1982). Nuc.Acids Res. 10, 4731-4751.

TOH, H., HAYASHIDA, H. and MIYATA, T. (1983). Nature 305, 827-829.

TOH, H., KIKUNO, R., HAYASHIDA, T., KUGIMIYA, W., INOUYE, S., YUKI, S. and SAIGO, K. (1985). EMBO J. 4, 1267-1272.

TRUETT, M.A., JONES, R.S. and POTTER, S.S. (1981). Cell 24, 753-763.

TWIGG, A.J. and SHERRATT, D.J. (1980). Nature 283, 216-218.

WEINER, A.M. (1980). Cell 22, 209-218.

WEINER, A.M., DEININGER, P.L. and EFSTRATIADIS, A. (1986). Ann.Rev. Biochem. 55, 631-661.

WILL, B.M., BAYEV, A.A. and FINNEGAN, D.J. (1981). J.Mol.Biol. 153, 897-915.

WILLIAMS, B.G. and BLATTNER, F.R. (1980). "Bacteriophage Lambda Vectors for DNA Cloning" in "Genetic Engineering", eds. Setlow, J.K. and Mullander, A. Vol.2, pp.201-281.

YOUNG, M.W. (1979). Proc.Natl.Acad.Sci. USA 76, 6274-6278.

Cell, Vol. 47, 1007–1015, December 26, 1986, Copyright © 1986 by Cell Press

Transposable Elements Controlling I–R Hybrid Dysgenesis in D. melanogaster Are Similar to Mammalian LINEs

D. H. Fawcett,* C. K. Lister,† E. Kellett, and D. J. Finnegan
Department of Molecular Biology
University of Edinburgh
Edinburgh EH9 3JR Scotland

Summary

I–R hybrid dysgenesis in D. melanogaster is controlled by transposable elements known as I factors. We have determined the base sequences of one complete I factor and the ends of six others. The ends of these elements are highly conserved and are flanked by target site duplications varying in length from 10–14 bp. There are no terminal repeats, and the 3′ end of one strand is A-rich, having 4–7 tandem repeats of the sequence TAA. This sequence organization is similar to that of mammalian LINEs, or L1 elements. The complete I factor sequence contains two long open reading frames, ORF1 and ORF2, of 1278 and 3258 bp. ORF1 encodes a possible nucleic acid–binding protein, and part of the amino acid sequence of ORF2 is similar to that of viral reverse transcriptases and polypeptides encoded by L1 elements. These results suggest that I factors transpose by reverse transcription of a full-length RNA.

Introduction

Hybrid dysgenesis is the production of abnormal characteristics in the progeny produced when particular strains of Drosophila melanogaster are crossed in an appropriate fashion. These traits include partial or complete sterility and increased frequencies of mutations and chromosome rearrangements. There are two independent systems of hybrid dysgenesis, P–M and I–R. P–M dysgenesis is produced by crossing M, maternal, strain females with P, paternal, strain males. In the I–R system R, reactive, strain females must be crossed with I, inducer, strain males. The progenies of the reciprocal crosses are apparently normal in each case (Kidwell, 1983; Bregliano and Kidwell, 1983).

The characteristics of P and I strains are controlled by transposable elements called P factors and I factors, respectively. These elements are activated when they are introduced into the cytoplasmic background of an M strain, in the case of P factors, or an R strain, in the case of I factors. This is believed to reflect the presence of regulatory molecules, which are produced by P or I factors and which are not present in the nuclei of M or R strains.

We have identified the I factor by analyzing the molecular lesions associated with eight *white* gene mutations, w^{IR1-8}, produced in I–R dysgenic females (Bucheton et al., 1984; Sang et al., 1984). Two of the mutations, w^{IR7} and w^{IR8}, determine a bleached white phenotype and are due to deletions of part of the *white* gene. The remaining six mutations determine colored eye phenotypes and have indistinguishable 5.4 kb elements inserted within the *white* gene. We believe that these 5.4 kb elements are I factors since several of these mutations are closely linked to I factor activity (Pelisson, 1981; Bucheton et al., 1984; Pelisson, personal communication) and complete 5.4 kb elements are only present in DNA from inducer strains (Bucheton et al., 1984).

We have now determined the base sequences of all of the I factor associated with the w^{IR1} mutation and of the ends of the I factors associated with mutations w^{IR2-6}. The structure of these elements is highly conserved and differs markedly from that of the P factor (O'Hare and Rubin, 1983) and most other transposable elements in D. melanogaster. The structure of the I factor does, however, resemble that of LINEs, or L1 elements (Singer, 1982; Singer and Skowronski, 1985), found in mammalian genomes. Like L1 elements, I factors encode a polypeptide with homology to reverse transcriptases.

Results

Sequence Analysis of Dysgenesis-Induced Mutations

The I factor associated with the w^{IR1} mutation is within a 6.2 kb SalI fragment (Figure 1; Bucheton et al., 1984). The complete base sequence of this fragment has been determined using the dideoxynucleotide method and is shown in Figure 2. The strategy for sequencing is shown in Figure 1 and is described in Experimental Procedures.

This fragment includes DNA from the *white* gene corresponding to nucleotides −1535 to −671 on the sequence of O'Hare et al. (1984). It differs from that reported for strain Canton S (O'Hare et al., 1984) by four base substitutions and two small insertion/deletions in addition to the 5.4 kb I factor insertion. The positions of these differences are marked on Figure 2. Only one of them, deletion of the G at position −859 in the sequence of O'Hare et al., would have any consequence for the coding capacity of the *white* gene. This would correspond to insertion of a C between bases 5942 and 5943 in Figure 2. The other differences either are in an intron or alter the third base of a codon without changing its sense. We have confirmed that the sequence of the wild-type *white* gene in the stock of strain Canton S cloned by Maniatis et al. (1978) is the same as that of w^{IR1} at position −859, suggesting that the proposed intron/exon structure of the *white* gene (O'Hare et al., 1984) may be incorrect in this region.

The I factor insertion is 5371 bp long and is flanked by direct repeats of a 12 bp target sequence (Figures 2 and 3). There is no sequence repeated at both ends of the I factor in either direct or inverted orientation. We have sequenced the ends of the I factors inserted into the *white*

* Present address: Institute of Animal Physiology and Genetic Research, Roslin, Midlothian EH25 9PS.
† Present address: John Innes Institute, Colney Lane, Norwich NR4 7UH.

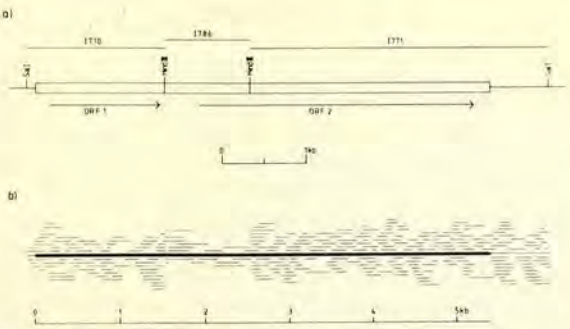


Figure 1. Map of the SalI Fragment from w^{IR1} Showing the Fragments Cloned for Sequencing

(a) Restriction map of the SalI fragment containing the I factor associated with the w^{IR1} mutation. The thin line represents *white* DNA and the box represents the I factor. The subclones from which fragments were generated for sequencing are shown above the map. The sequencing method is described in Experimental Procedures. The sequence across the left-hand HindIII site within the I factor was determined by sequencing this region of a partial HindIII digestion product cloned from w^{IR1} (Bucheton et al., 1984). The sequence across the right-hand HindIII site was determined by sequencing this region of the complete I factor cloned from w^{IR3} (Bucheton et al., 1984). The positions of the two long open reading frames encoded by the I factor are shown below the map.

(b) The long horizontal line represents the SalI fragment shown in (a), with the thicker part being the I factor. The short lines above and below the horizontal represent regions cloned and sequenced in M13 vectors. Regions sequenced in the upper strand (left to right, 5′ to 3′) are shown above the horizontal, while regions sequenced in the lower strand (right to left, 5′ to 3′) are shown below.

gene on chromosomes carrying mutations w^{IR2-6}, and the results are shown in Figure 3. Each insertion is flanked by a target site duplication, although their lengths are slightly variable. The I factors associated with mutations w^{IR1}, w^{IR3}, and w^{IR4} are inserted at exactly the same position and lie within the fifth intron of the *white* gene (O'Hare et al., 1984). The 14 bp duplication flanking the w^{IR3} insertion contains an extra A with respect to the w^+ (Canton S), w^{IR1}, and w^{IR4} alleles. This extra base may have been present in the parental w^+ chromosome before insertion, or it may have been inserted during I factor transposition.

The I factor associated with the w^{IR5} mutation lies within the 3′-untranslated region of the *white* gene. The insertion associated with the w^{IR6} mutation is within the first intron. The sequence organization adjacent to the left-hand end of this element is complex. The last 3 bases of the target site duplication, TAACAACCAG, are the same as the first 3 bases of the I factor. They form part of a tandem duplication of the first 6 bases of the I factor, lying adjacent to the insertion and separated from it by the sequence AAT (Figure 3). We presume that this structure was formed by the enzymes responsible for generating the target site duplication during transposition.

The w^{IR2} mutation arose on a chromosome that already carried the *white* gene mutation w^1 (Sang et al., 1984). This mutation is associated with an insertion of an F-like element (Zachar and Bingham, 1982; O'Hare et al., 1983). The w^{IR2} I factor lies within the w^1 element and is flanked by a 12 bp direct repeat. We have determined the sequence of the corresponding region of the w^1 insertion from a clone supplied by K. O'Hare. This indicates that the 12 bp repeat is a target site duplication.

Peifer and Bender (1986) have found a spontaneous mutation of the bithorax complex, bx^{F31}, which is associated with insertion of what appears to be a complete I factor. They have kindly sent us a clone of this insertion, and we have determined the sequence of its ends (Figure 3). It is flanked by a direct repeat of 13 bp. We presume that this is a target site duplication, although we do not know the corresponding wild-type sequence.

The ends of the seven I factors that we have studied are highly conserved (Figure 3). We have sequenced the first 110 bp of the elements associated with mutations w^{IR1-5} and have less extensive information for w^{IR6} and bx^{F31}. Only one position varies—the third base, which is T in the case of w^{IR1} and w^{IR2}, and G in the case of w^{IR3-6} and bx^{F31}. The last 210 bp of the w^{IR1-5} insertions are identical except for the number of tandem repeats of the sequence TAA. Again, we have less information for w^{IR6} and bx^{F31}, but the pattern is the same. We have found elements with 4–7 TAAs. The number cannot be determined precisely for w^{IR1}, w^{IR6}, and bx^{F31} since, in each case, one copy could be part of the target site duplication (Figures 2 and 3).

The two remaining *white* mutations, w^{IR7} and w^{IR8}, are associated with deletions (Sang et al., 1984). We have cloned DNA including the breakpoints of these deletions, and have determined the base sequences surrounding them (Figure 4). The length of the w^{IR7} deletion is not known. Its distal breakpoint is between bases 5630 and 5631 in Figure 2, while its proximal breakpoint is outside the *white* locus and may be proximal to the gene *roughest* (Sang et al., 1984). The w^{IR8} deletion is about 2.4 kb long. Its proximal breakpoint is between bases 5601 and 5602 in Figure 2, and its distal breakpoint is about 400 bp distal to the SalI site at position −3050 in the sequence of O'Hare et al. (1984). The sequences spanning the breakpoints of these deletions are shown in Figure 4. Neither of these alleles has any I factor sequences at the site of the deletion, so we cannot say whether they resulted directly from I–R dysgenesis.

Coding Sequences within the I Factor

We have found the positions of methionine and chain-terminating codons in each of the six open reading frames of the I factor (Figure 5). There are only two long open reading frames, both going from left to right in Figure 1. Their amino acid sequences are shown in Figure 2. They are 1278 bp and 3258 bp long and are separated by 471 bp. The first, ORF1, starts 187 bp from the left-hand end of the I factor; the second, ORF2, ends 179 bp before the right-hand end.

Kozak (1986) has found that initiating ATG codons are preferentially included in sequences having the consensus ACCATGG, with the A at position −3 and the G at position +4 being most strongly conserved. The first ATG in ORF1 is the fourth codon and is included in the sequence ATCATGA, suggesting that it is likely to be used for initiation. ORF2 has ATG as codons 2 and 25, and

```
(-1535)
-99   gtcgacttcgggcctccctcataaaaactggcagctctgaggtgaacacctaaatcgaatcgattcattagaaagttagtaaattattaatatgcaaagt   1
2     ATTACCACTTCAACCTCCGAAGAGATAAGTCGTGCCTCTCAGTCTAAAGCCTCGCTTCGCGTAAGCCCAAAACTCTTATCAGCAAAATCTTGATAAACAA   101
102   ATATCAACCACAAAGAGAAAATAAAAAACTTAACAACAAAAACAACAATACCGCTAATCCGGGCTCAAGCCCTTAACCAACAATCATGACAGACCCACCA   201
                                                                                                  ProThrIleMetThrAspProPro
202   AACATTTACAAAATCACTTCAAAAACATACCAATCCCAATTAGGCGAACCTAAATTTATAATTATTAAAAGAAATGACAACAACTCTTTCGAAAGAACTT   301
      AsnIleTyrLysIleThrSerLysThrTyrGlnSerGlnLeuGlyGluProLysPheIleIleIleLysArgAsnAspAsnAsnSerPheGluArgThrS
302   CACCATTCATCATAAAAAAATCGGTGGACTTTGCCTGTGGAGGAGAAGTTGAGGGATGCAAACGTACAAGAGACGGCAACCTGCTAATAAAAACCAAAAA   401
      erProPheIleIleLysLysSerValAspPheAlaCysGlyGlyGluValGluGlyCysLysArgThrArgAspGlyAsnLeuLeuIleLysThrLysAs
402   TGAATTACAAGCCAGAAAACTCCTAAAACTAACAAAAATTGCAGATGAGGATGTAACAGCAAGTGAACATAAAACATTAAACTTCTCTAAGGGAGTTATT   501
      nGluLeuGlnAlaArgLysLeuLeuLysLeuThrLysIleAlaAspGluAspValThrAlaSerGluHisLysThrLeuAsnPheSerLysGlyValIle
502   TACTGTAACGACCTTAGACACATCGACGAAGACACAATTCTACAAGAACTAAAACCACAAAAAGTATCTGAAGTTAAAAAAATAATGAAACGGCAAAACC   601
      TyrCysAsnAspLeuArgHisIleAspGluAspThrIleLeuGlnGluLeuLysProGlnLysValSerGluValLysLysIleMetLysArgGlnAsnP
602   CCAACTCTAACTCCGACACCAACAACATCACATTAGTTGAAACTGGACTCATAATTATAACCTTTGAATCGCATAAGCTCCCCGAGATAGTACGAATCGG   701
      roAsnSerAsnSerAspThrAsnAsnIleThrLeuValGluThrGlyLeuIleIleIleThrPheGluSerHisLysLeuProGluIleValArgIleGI
702   GTACGAAACAGTCCGAGTACGAGACTATATCCCACTCCCACTTCGATGCAAAAATGCCTCCGCTTCGGTCATCCAACACCCATATGCAAAGTGTAGAAA   801
      yTyrGluThrValArgValArgAspTyrIleProLeuProLeuArgCysLysLysCysLeuArgPheGlyHisProThrProIleCysLysSerValGlu
802   ACTTGCATCAATTGCTCTGAAACAAAACACACAAACGACGGAGAAAAATGCACAAACGAAAAAAACTGCTTAAATTGCCGAAATAACCCAGAACTTGACC   901
      ThrCysIleAsnCysSerGluThrLysHisThrAsnAspGlyGluLysCysThrAsnGluLysAsnCysLeuAsnCysArgAsnAsnProGluLeuAspH
902   ATCAACACAGCCCAATTGACCGCAAATGCCCTACGTTCATAAAAAACCAGGAATTAACAGCAATTAAAACCACACAAAAAGTTGACCATAAAACGGCCCA   1001
      isGlnHisSerProIleAspArgLysCysProThrPheIleLysAsnGlnGluLeuThrAlaIleLysThrThrGlnLysValAspHisLysThrAlaGI
1002  ACACATATATTTCGAACGTCACGGCTTCCAAACGAAAAACACCTACGCCAAAACACTTACAAACGGCACAACCCAGAGGACAACAAACACTCCATCACCT   1101
      nHisIleTyrPheGluArgHisGlyPheGlnThrLysAsnThrTyrAlaLysThrLeuThrAsnGlyThrThrGlnArgThrThrAsnThrProSerPro
1102  AATATTCACACAAACACAACCCAATCACAACAACAAAATCCGCACCACACACCCAAATCAGCAGCACAAAACACTTCAGCTAAGACACCAACAACTGAAC   1201
      AsnIleHisThrAsnThrThrGlnSerGlnGlnGlnAsnProHisHisThrProLysSerAlaAlaGlnAsnThrSerAlaLysThrProThrThrGluP
1202  CAGCCAAAACAACCTTACTATCCAACCAACCACACCAACACCACCACCACCACAGCTACGACAAACTAGAAGACATGGATACCGACTACACACCTACCAG   1301
      roAlaLysThrThrLeuLeuSerAsnGlnProHisGlnHisHisHisHisHisSerTyrAspLysLeuGluAspMetAspThrAspTyrThrProThrAr
1302  AAAACCATCTACGACATACTCATCACAACTCACAGAAGACCTAAAAATAAAAATCTTCCCTAAAGATAAGTCCAATAACCTATCCATAAACCTTAAAGCA   1401
      gLysProSerThrThrTyrSerSerGlnLeuThrGluAspLeuLysIleLysIlePheProLysAspLysSerAsnAsnLeuSerIleAsnLeuLysAla
1402  TCAAAACTAAAGGCCAAAGCCCACAAAAACAAGCACACTAACAACAGCGACAGCGAATCCATATAGAACTCTACACAAAACCCTAACCGTTAACACTACC   1501
      SerLysLeuLysAlaLysAlaHisLysAsnLysHisThrAsnAsnSerAspSerGluSerIle***
1502  TTTAAGTAAGTTATAAGCTTTAATTTTCTCACAAATGTCCCTAACTATAATCCAATGGAATCTAAAAGGATATCTAAACAACTACAGCCATCTCCTTATT   1601
1602  CTAATCAAAAAATACTCCCCCCACATAATTTCCCTCCAAGAAACCCATATACAATACACTAATAACATTCCAACCCCAATAAACTACAAACTATTAACAA   1701
1702  ATATTGCCACCAACAGATTTGGGGGGCGTACGACTACTAGTGCATAAGTCAATACAACACACTGTCCTCAACATAACAATCGATATAGAAGCAATAGCCA   1801
1802  TAAATATAGAATCTAAACTTAAATTAAACATATTTTCCACATACATTTCTCCGACCAAAAACATAACTAACCAGACACTCCATAACACATTTAACATACA   1901
1902  ACAAACACCCTCTCTAATTACGGGAGATTTAATGGATGGCACCATCCTGGGGCTCCCCAACAACAAATAAACGAGGAAAAAATAACTCATAGATTCATTG   2001
                                                                                                  TrpMetAlaProSerTrpGlySerProThrThrAsnLysArgGlyLysIleThrHisArgPheIleA
2002  ACAACATGCACCTTATCCTGTTAAACGACAAATCTCCCCACACACTTTTCAACACACAATACATACACACACATAGACCTCACACTCTGCTCTCCAATCCT   2101
      spAsnMetHisLeuIleLeuLeuAsnAspLysSerProThrHisPheSerThrHisAsnThrTyrThrHisIleAspLeuThrLeuCysSerProIleLe
2102  AGCCCCCCACGCCAAGTGGAAAATACTAAACGATCTTCACGGTAGCGACCATTTCCCTATTATCACAACACTATTCCCAACAACCAATCCACAAAAATTC   2201
      uAlaProHisAlaLysTrpLysIleLeuAsnAspLeuHisGlySerAspHisPheProIleIleThrThrLeuPheProThrThrAsnProGlnLysPhe
2202  TACAGACCCTTTTTTAAACTCAAAGAAGCCAACTGGGAACAGTTCAACGCTCTTACCCACCAAACCAACAAGAAATACCCCACCTCCCACAACGTAAACA   2301
      TyrArgProPhePheLysLeuLysGluAlaAsnTrpGluGlnPheAsnAlaLeuThrHisGlnThrAsnLysTyrProThrSerHisAsnValAsnL
2302  AAGAAGCCGCTCTAATCAATAGAATCATCCTTTATAGCGCAAACCTCTCCATCCCACAAACCTCACCTAACACACATCCATACAGGGTTCCATGGTGGAA   2401
      ysGluAlaAlaLeuIleAsnArgIleIleLeuTyrSerAlaAsnLeuSerIleProGlnThrSerProAsnThrHisProTyrArgValProTrpTrpAs
2402  TAAACACCTCGACCAATTACGTAAAGAAAAACAACTTGCCTGGAAAAAATTAAACCGCACAATTACTGTTGACAACATTCTAGACTATAGACGCAAAAAC   2501
      nLysHisLeuAspGlnLeuArgLysGluLysGlnLeuAlaTrpLysLysLeuAsnArgThrIleThrValAspAsnIleLeuAspTyrArgArgLysAsn
2502  GCAATATTTAGATACGAACTAAAAAAGAGGAAAAAAGAAGCTTCCAGCTCTTTCACCTCAACCATCCATCCCACTACTCCCTCATCCAAAATATGGGCCA   2601
      AlaIlePheArgTyrGluLeuLysLysArgLysLysGluAlaSerSerSerPheThrSerThrIleHisProThrThrProSerSerLysIleTrpAlaA
2602  ATATAAGACGCTTCTGCGGACTTAACCCAGCAAAACAAATTCATGCCATCACAAACCCAGTAAATAACGAGACTACATTGGCTAGCAACGAAATTGCTAA   2701
      snIleArgArgPheCysGlyLeuAsnProAlaLysGlnIleHisAlaIleThrAsnProValAsnAsnGluThrThrLeuAlaSerAsnGluIleAlaAs
2702  CATATTCGCACAACATTTCTCTGACCTCTCCGGCGACTGGAACTTCTCAGAGGAGTTCCGGAACAATAAATATAGAAATAACATACATCTCTACACCCCC   2801
      nIlePheAlaGlnHisPheSerAspLeuSerGlyAspTrpAsnPheSerGluGluPheArgAsnAsnLysTyrArgAsnAsnIleHisLeuTyrThrPro
2802  TCTCCAATAGCCCAAACCATAGAAGAGAACATAACGTATCTAGAACTTAGCTCAGCACTACAAACATTAAAAGGATGTGCTCCAGGACTAAATAGAATCT   2901
      SerProIleAlaGlnIleArgIleGluAsnIleThrTyrLeuGluSerSerAlaLeuGlnThrLeuLysGlyCysAlaProGlyLeuAsnArgIleS
2902  CGTATCAAATGATCAAAAATAGCTCCCACACAACAAAAAACCGAATAACGAAACTATTTAATGAAATATTCAATAGCCACATACCTCAAGCCTACAAAAC   3001
      erTyrGlnMetIleLysAsnSerSerHisThrThrLysAsnArgIleThrLysLeuPheAsnGluIlePheAsnSerHisIleProGlnAlaTyrLysTh
```

```
3002  AAGCCTAATCATCCCAATCCTTAAGCCAAACACCGACAAAACGAAAACTTCCTCATACCGACCCATCTCCCTCAACTGCTGTATAGCAAAGATACTTGAT   3101
      rSerLeuIleIleProIleLeuLysProAsnThrAspLysThrLysThrSerSerTyrArgProIleSerLeuAsnCysCysIleAlaLysIleLeuAsp
3102  AAAATAATTGCGAAAAGACTCTGGTGGCTAGTGACATATAACAACCTAATTAACGACAAACAATTCGGGTTCAAAAAAGGCAAATCGACTTCGGACTGTC   3201
      LysIleIleAlaLysArgLeuTrpTrpLeuValThrTyrAsnAsnLeuIleAsnAspLysGlnPheGlyPheLysLysGlyLysSerThrSerAspCysL
3202  TACTCTATGTAGACTATCTCATAACGAAGTCAAAAATGCACACACCTCCCTCGTCACTCTTGATTTTTCAAGAGCCTTCGATCGAGTAGGTGTGCACTCCAT   3301
      euLeuTyrValAspTyrLeuIleThrLysSerLysMetHisThrSerLeuValThrLeuAspPheSerArgAlaPheAspArgValGlyValHisSerIl
3302  AATCCAGCAATTGCAGGAATGGAAAACGGGTCCCAAAATAATAAAATACATTAAAAACTTCATGAGCAACAGAAAAATAACTGTCCGCGTCGGTCCGCAT   3401
      elleGlnGlnLeuGlnGluTrpLysThrGlyProLysIleIleLysTyrIleLysAsnPheMetSerAsnArgLysIleThrValArgValGlyProHis
3402  ACATCAAGCCCGTTACCCCTATTCAACGGAATCCCCCAAGGTTCACCCATATCCGTAATACTTTTCCTCATAGCATTCAACAAATTATCCAACATCATAT   3501
      ThrSerSerProLeuProLeuPheAsnGlyIleProGlnGlySerProIleSerValIleLeuPheLeuIleAlaPheAsnLysLeuSerAsnIleIleS
3502  CCCTACATAAAGAAATTAAATTCAACGCATATGCCGACGACTTCTTCCTTATAATAAATTTCAACAAAAACACAAATACAAATTTCAACTTAGACAATCT   3601
      erLeuHisLysGluIleLysPheAsnAlaTyrAlaAspAspPhePheLeuIleAsnPheAsnLysAsnThrAsnThrAsnPheAsnLeuAspAsnLe
3602  ATTCGACGATATAGAAAATTGGTGCTCCTACTCAGGGGCATCGCTTTCCCTATCCAAATGTCAACACCTCCACATATGCAGAAAACGTCACTGCACATGC   3701
      uPheAspAspIleGluAsnTrpCysSerTyrSerGlyAlaSerLeuSerLeuSerLysCysGlnHisLeuHisIleCysArgLysArgHisCysThrCys
3702  AAGATAAGCTGCAACAACTTCCAAATTCCTAGCGTTACGTCCTTAAAAATTCTAGGAATAACCTTAAACAACAAATACAAATGGAACACACACATAAACC   3801
      LysIleSerCysAsnAsnPheGlnIleProSerValThrSerLeuLysIleLeuGlyIleThrAsnAsnLysTyrLysTrpAsnThrHisIleAsnL
3802  TACTTCTACCCAAACTACACAACAAGCTAAATATAATAAAATGCCTATCTAGTCTTAAATTTAACTGCAACACGCATACACTACTTAATGTCGCAAAAGC   3901
      euLeuLeuProLysIleHisAsnLysLeuAsnIleIleLysCysLeuSerSerLeuLysPheAsnCysAsnThrHisThrLeuLeuAsnValAlaLysAl
3902  AACAATTATAGCCAAACTAGAGTATGGTTTGTTTCTGTACGGCCATGCTCCCAAAAGCATTTTAAACAAAATAAAAACACCGTTTAACTCCGCTATCCGT   4001
      aThrIleIleAlaLysLeuGluTyrGlyLeuPheLeuTyrGlyHisAlaProLysSerIleLeuAsnLysIleLysThrProProHeAsnSerAlaIleArg
4002  CTAGCTCTCGGCGCATATCGCTCTACCCCAATAAATAACTTACTTTACGAATCGAATACTCCCCCCTTAGAAATGAAACGAGACCTTCAAATAGCCAAAC   4101
      LeuAlaLeuGlyAlaTyrArgSerThrProIleAsnAsnLeuLeuTyrGluSerAsnThrProProLeuGluMetLysArgAspLeuGlnIleAlaLysL
4102  TATCCCAAAACCTAATCCTCTCCAAAAACACACCAATACATAAGTTCTTAAAGCCTAAAAAAGCTAATAAGAAAAAAACATCAACAATAGACCGAACAAT   4201
      euSerGlnAsnLeuIleLeuSerLysAsnThrProIleHisLysProLeuLysLysAlaAsnLysLysLysThrSerThrIleAspArgThrIl
4202  CAAACTTAGCCTAGAACTTAATCTACCCTACAAACCAATAAAACTCCATAAAAACAAACCACCATGGACCCTCCCCAATCTAATAGACACGTCACTTAGA   4301
      eLysLeuSerLeuGluLeuAsnLeuProTyrLysProIleLysLeuHisLysAsnLysProProTrpThrLeuProAsnLeuIleAspThrSerLeuArg
4302  ATCCATAAGAAAGAACAAACATCTCCAGACCAATACAGAAAAATTATACGAACACACAAAGAATAACCTCAAAACACACAATTTCATATTCACTGACGGTT   4401
      IleHisLysLysGluGlnThrSerProAspGlnTyrArgLysLeuTyrGluHisThrLysAsnAsnLeuLysThrHisAsnPheIlePheThrAspGlyS
4402  CAAAAATTAATTACACAATATCATTCGCCATTACAACGGAGACAGACGTCTTGAAATACGGCATACTGCCCCCATATTCATCCGTCCTCACCTCCGAAAC   4501
      erLysIleAsnTyrThrIleSerPheAlaIleThrThrGluThrAspValLeuLysTyrGlyIleLeuProProTyrSerSerValLeuThrSerGluTh
4502  AATCGCCATCCTAGAAGCAATAGAACTTACTAAAAACCGAAGAGGCAAATTTATTATCTGCTCCGACTCCCTATCAGCAGTAGATTCAATTCAAAACACA   4601
      rIleAlaIleLeuGluAlaIleGluLeuThrLysAsnArgArgGlyLysPheIleLeuCysSerAspSerLeuSerAlaValAspSerIleGlnAsnThr
4602  AATAATAACAGCTTTACCCAAGCAGAATACGATCGCTAATAACGCAACACGCACCTAAAATTAAAATAATGTGGATTCCTGGCCATTCAGGAATAAAAG   4701
      AsnAsnAsnSerPheTyrProSerArgIleArgSerLeuIleThrGlnHisAlaProLysIleLysIleMetTrpIleProGlyHisSerGlyIleLysG
4702  GAAATGAATTAGCCGATCAAGCTGCAAAATCAGCAAGCAGTATGCCACCTTATCCTCACCCCAAACATAAATACCACAGATATAAAAAAACACCTTAAAGC   4801
      lyAsnGluLeuAlaAspGlnAlaAlaLysSerAlaSerSerMetProLeuLeuLeuThrProAsnIleLysLysIleLeuLysAl
4802  CGACCTTGCGACAAAACAGAAAGAACACATAATAAACTGCAGTCCATGGTACCAATCTATTAACACGAACACCTCACACCCATGCGATTACCTTAAACAA   4901
      aAspLeuAlaThrLysGlnLysGluHisIleIleAsnCysSerProTrpTyrGlnSerIleAsnThrAsnThrSerHisProCysAspTyrLeuLysGln
4902  TCCCACCCAAATTGGACCAGACTCGACCAAATAAAAATAATACGACTTCGACTAGGACACACAAACATAACCCACCAACACTACCTAAATCCCAATTCAA   5001
      SerHisProAsnTrpThrArgLeuAspGlnIleLeuLysIleIleArgLeuArgLeuGlyHisThrAsnIleLeuSerIleAspSerLeuSerAlaVal
5002  TACCAACTTGCCCGTTTTGCCAAGGTGATATTTCTTTAAACCACATATTTAACTCATGCCCATCCCTCCTACAAACCAAGCAAGATATATTTAACAACAC   5101
      IleProThrCysProPheCysGlnGlyAspIleSerLeuAsnHisIlePheAsnSerCysProSerLeuLeuGlnThrLysGlnAspIlePheAsnAsnTh
5102  CAACCCTCTAGACCTTCTTAGCAAACCCAATCCAGATAACATACAAAAACTCATACTTTTCCTCAAAAAAACTAAATTATACCACAAAATCTAAAAACAA   5201
      rAsnProLeuAspLeuLeuSerLysProAsnProAspAsnIleGlnLysLeuIleLeuPheLeuLysLysThrLysLeuTyrHisLysIle***
5202  AACAGGCATTTGTACATAACAAGCCAGCAATTAGTTACCAAATTAGATATTAACTAAATTAAGATATAATAACATTGTAAATAAATATAGCTGTAAGCCC   5301
5302  CGTAGCTAATGCTATACTATCTAAGTTAGTCTAGTTTTGTAAACTATTCTATCTATCATAATAATAATAAtatgcaaagtgtattctaaacaagactt   5401
5402  acatttatcgtggcaaagacgttttgaaaggtcatgttggtcaggaagaggaagatggctccgttgatattcatcacgcccactttgcgtgagttgttggc   5501
5502  ccaaaaagatgaggccaatcaagatggcaaccatctgcaaattaaaatgttactcgcatctcattaatattcgcgagttaaatgaaatttatttatcttc   5601
5602  tgcaaaactataaactatacatctcattgaaaaaaactaagaagggtgtggaatcaggcaattctatctaaaatctagcgaatttgtttccaagaattgt   5701
5702  aagcgttatatcatttgtttccactggaaccactcaccgttgtctgaataagtcgcactttacgaggagtggttccttgagcaccgacagccaggatcg   5801
5802  ccacaggaccgcccggaactgcatgaaccaggtggccttgtaggtgtacccattctccggctgctccagtggcttctccagattttttggtggccaacaac   5901
5902  tgctccatatcccgggctactttgctaatggcaaaattgtcgcatatcttggcgatccgatcacgggactcgatctcccgtccgggcacaacggccaaca   6001
6002  cctgtacgtaaaagtccgccggattgtagttggtaggacactgggcacccacgctggatagggattgagatgtaatgctagatacccctaataaa   6101
6102  cacatcgaactcactaggaaaagaagtcgac   6132
                                        (-671)
```

Figure 2. Sequence of the I Factor from the *w*^{R1} Mutation

This sequence is that of the Sall fragment shown in Figure 1. The sequence of the I factor is shown in upper case; that of adjacent *white* DNA, in lower case. The *white* sequence is the reverse complement of bases −671 to −1535 of O'Hare et al. (1984). The differences between the two sequences are shown. The number 17 indicates an insertion/deletion of 17 bases. This is the reverse complement of the published sequence. The target site duplication is boxed. The dotted line indicates that the first 3 bases, TAA, could be either part of this duplication or part of the I factor. The amino acid sequences of the two long open reading frames in the I factor are shown below the base sequence.

LEFT-HAND END	TARGET SITE DUPLICATION		RIGHT-HAND END
tattaatatgcaaatCATTAC	w^{IR1}	12/9	TCA(TAA)$_4$taatatgcaaatgta
attaatatgcaaaatCAGTAC	w^{IR3}	14	TCA(TAA)$_5$ttaatatgcaaaatg
tattaatatgcaaatCAGTAC	w^{IR4}	13	TCA(TAA)$_5$ttaatatgcaaatgt
ttatttactgcagagCATTAC	w^{IR2}	12	TCA(TAA)$_7$ttactgcagagttt
atatccgaaataactCAGTAC	w^{IR5}	12	TCA(TAA)$_6$tccgaaataactgct
tataaaaggccgaaaCAGTAC	bx^{F31}	13/10	TCA(TAA)$_6$taaaaggccgaaacc
atataaacaaccagtaccagtacaatCAGTAC	w^{IR6}	10/7	TCA(TAA)$_6$taacaaccagatatt

Figure 3. Sequences at the Ends of I Factor Insertions

The sequences of the ends of the I factors are shown in upper case, and the sequences of adjacent DNA are shown in lower case. The target site duplications associated with each I factor are boxed. The duplications associated with w^{IR1}, w^{IR6}, and bx^{F31} cannot be defined precisely since, in each case, one copy of the TAA repeat could be part of either the duplication or the I factor. This is indicated by a dotted line. The braces indicate tandem repeats of the first 6 bases of the I factor, which lie adjacent to the w^{IR6} insertion. Cloned fragments of genomic DNA used to obtain these data are described in the text, except for w^{IR3} cloned by Bucheton et al. (1984), w^{IR2} and w^{IR5} cloned by Sang et al. (1984), and w^{IR4} and w^{IR6} cloned as described in Experimental Procedures.

w^{IR7}	ATAAACTATACATCTCATTGTATATATGTATATATATATA
w^{IR8}	TGTATGTTTGTATATATAAATGCAAAACTATAAACTATAC

Figure 4. Sequences at the Deletion Breakpoints of w^{IR7} and w^{IR8}

The vertical line indicates the deletion breakpoints. The base immediately before the breakpoint in w^{IR7} is 5630 in Figure 2. The base immediately following the breakpoint in w^{IR8} is 5602 in Figure 2. The w^{IR7} and w^{IR8} fragments were cloned as described in Experimental Procedures.

these are included in the sequences TGGATGG and AACATGC, respectively. Ninety-five percent of natural initiation codons have A at position −3 (Kozak, 1984), suggesting that the ATG at position 25 is the one more likely to be used. The predicted relative molecular masses of polypeptides initiated from these positions are 48 kd and 121 kd, respectively. Within ORF1, there are no obvious donor splice sites that could be used to join parts of the two open reading frames, but we cannot exclude this.

Comparison with Other Transposable Elements

The structure of the I factor differs from that of most other transposable elements in that it does not have terminal repeat sequences (Finnegan, 1985). The elements that most closely resemble the I factor are the long interspersed sequences, or LINEs, found in mammalian genomes (Singer, 1982). Each studied species has a single major family of LINEs, known as L1 elements. They are repeated 10^4 to 10^5 times per genome, and make up several percent of the total DNA. Complete L1 elements are 6–7 kb long and have an A-rich sequence at the 3′ end of one strand, called the 3′ end of the element, often

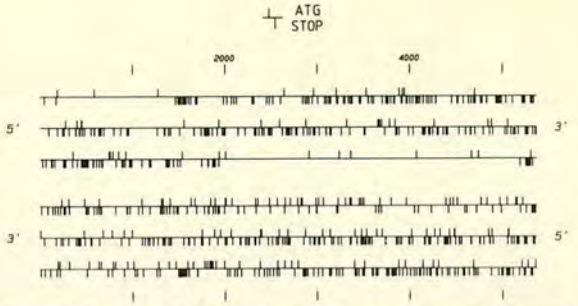


Figure 5. Positions of ATG and Stop Codons within the I Factor from w^{IR1}

The six possible reading frames are represented by horizontal lines. Vertical lines above the horizontal indicate ATG codons. Vertical lines below the horizontal indicate translational stop codons.

preceded by the polyadenylation signal AATAAA. Many copies of the L1 element are truncated at the other end (the 5′ end).

This sequence organization is similar to that of processed pseudogenes (Rogers, 1985) and other retroposons (Rogers, 1983, 1985). L1 elements are often flanked by short direct repeats, which in some cases are known to be target site duplications (Gebhardt and Zachau, 1983; Singer and Skowronski, 1985), and for this reason L1 elements are thought to be transposable. Jagadeeswaran et al. (1981) and Van Arsdell et al. (1981) suggested that short retroposons, or SINEs (Singer, 1982), might transpose by reverse transcription of an RNA intermediate. This model has also been applied to LINEs (Singer and Skowronski, 1985; Rogers, 1985).

The genetic properties of I factors indicate that they determine a function that is required for transposition and a function that regulates transposition (Bregliano and Kidwell, 1983). Since the structure of the I factor resembles that of LINEs, we speculated that they might transpose by a similar mechanism and wondered whether one of the open reading frames in the I factor might encode a reverse transcriptase. Several groups have compared the amino acid sequences of retroviral reverse transcriptases and have found that they share short regions of highly conserved amino acid sequences (Toh et al., 1983; Patarca and Heseltine, 1984; Toh et al., 1985). Toh et al. (1983) have shown that amino acids corresponding to these conserved residues are also present in the putative polymerase genes of human hepatitis B virus, HBV, and cauliflower mosaic virus, CaMV (Figure 6), both of which are thought to replicate by reverse tanscription. The same is true of the transposable elements *copia* and the *copia*-like element *17.6*, of D. melanogaster (Saigo et al., 1984; Emori et al., 1985; Mount and Rubin, 1985) and the *Ty* elements of S. cerevisiae (Clare and Farabaugh, 1985). These probably all transpose by a mechanism related to a retroviral life cycle and involving reverse transcriptase (Boeke et al., 1985).

We have found no homology between ORF1 of the I factor and reverse transcriptases, but ORF2 has regions very similar to the conserved domains mentioned above (Figure 6). Toh et al. (1983) found 10 invariant amino acids

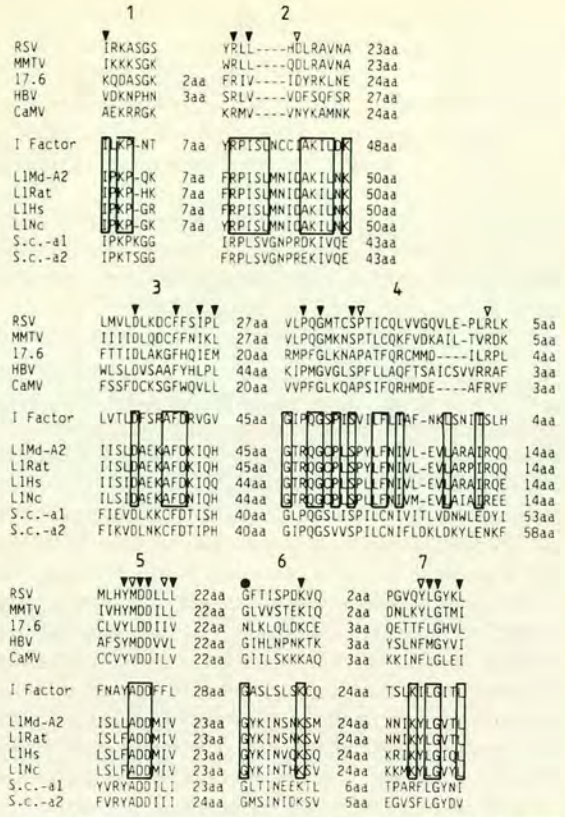


Figure 6. Comparison of ORF2 of the I Factor with Known and Probable Reverse Transcriptases

Seven conserved regions of amino acid sequence (Toh et al., 1985; Hattori et al., 1986) are shown in segments 1–7. The viral sequences are as follows: RSV (Schwartz et al., 1983); MMTV, murine mammary tumor virus (Chiu et al., 1985); HBV (Galibert et al., 1979); CaMV (Gardner et al., 1981). *17.6* is sequence from the *copia*-like element *17.6* (Saigo et al., 1984). I factor indicates sequence from ORF2 of the I factor shown in Figure 2. The L1 element sequences are, L1Md-A2, mouse L1 element A2 (Loeb et al., 1986); L1Rat, rat L1 element LINE 3 (D'Ambrosio et al., 1986); L1Hs, consensus human L1 sequence; L1Nc, consensus Nycticelus coucang sequence (Hattori et al., 1986). S.c.-a1 and S.c-a2 are sequences encoded by introns a1 and a2 of the mitochondrial cytochrome oxidase subunit I gene of S. cerevisiae (Bonitz et al., 1980). The triangles symbols indicate positions at which Toh et al. (1985) found identical or similar amino acid residues when comparing the sequences of eight enzymes known, or believed, to be reverse transcriptases. A filled triangle indicates that the sequence of the I factor codes for an identical or similar residue at this position. An open triangle indicates that the I factor codes for an unrelated residue at this position. The filled circle indicates a position at which five viral reverse transcriptases and the I factor code for the same residue. The boxes indicate positions at which the I factor and each of the L1 elements code for identical residues. The numbers indicate the number of residues separating each block of sequence. Amino acid abbreviations follow the standard single letter code. The following groups of residues with similar properties have been used for comparing sequences: P, A, G, S, and T (neutral or weakly hydrophobic); Q, N, E, and D (hydrophylic, acid amine); H, K, and R (hydrophylic, basic); L, I, V, and M (hydrophobic); F, Y, and W (hydrophobic, aromatic); C (cross-link forming).

when comparing the putative RNA-dependent DNA polymerases of Moloney murine leukemia virus (M-MuLV), Rous sarcoma virus (RSV), CaMV, HBV, and woodchuck hepatitis virus (WHV). Eight of these are present in ORF2 (Figure 6). When Toh et al. (1985) increased their set of putative reverse transcriptases to include those encoded by the *copia*-like element *17.6*, human T cell leukemia virus I (HTLVI), and duck hepatitis virus (DHBV), they found 27 positions containing identical or chemically similar amino acid residues. The sequence encoded by ORF2 contains amino acids corresponding to 21 of these (Figure 6, three of these positions are not shown in this figure). Since ORF2 has sequences corresponding to all of the highly conserved regions of reverse transcriptases, we suggest that it may also encode an RNA-dependent DNA polymerase.

The complete sequence of apparently full-length L1 elements have been determined for mouse (Loeb et al., 1986) and rat (D'Ambrosio et al., 1986), and consensus sequences have been constructed for L1 elements from human and prosimian genomes (Hattori et al., 1986). The L1 element from mouse, L1Md-A2, contains open reading frames of 1137 bp and 3900 bp, the second of which contains regions similar to reverse transcriptases (Figure 6). Similar regions of homology can be found in open reading frames encoded by the rat L1 and the human and prosimian consensus sequences (Figure 6).

In view of the structural similarities between the I factor and L1 elements, and because both classes of element encode polypeptides with some homology to reverse transcriptases, we have compared the second open reading frames of the I factor and L1Md-A2 directly (Figure 6). The strongest homology is in region 2 of Figure 6 where 10/16 residues are identical between the I factor and L1Md-A2. In contrast, only 1/16 residues are identical between the I factor and RSV, and 2/16 between L1Md-A2 and RSV. The corresponding figures for amino acid residues that are either identical or similar in this region are 12/16, 6/16, and 8/16. The rat L1 element and the consensus sequences for human and prosimian elements show a similar degree of homology to ORF2 (Figure 6). The number of amino acid residues separating these regions of homology is also very similar in the case of the I factor and L1 sequences. There are no regions of extensive homology between the I factor and L1 sequences outside those shown in Figure 6.

Burton et al. (1986) have compared cloned mouse and human L1 sequences by Southern transfer experiments and have detected two regions of conserved sequence, CS1 and CS2, that cross-hybridize more strongly than others. They have extended this analysis to include a wide variety of other mammals, probing digests of genomic DNA with cloned mouse L1 sequences. Again probes including CS1 and CS2 showed the strongest cross-hybridization. Interestingly the region of L1 elements that is most similar to the I factor (region 2 in Figure 6) is included within CS2, suggesting that this region is subject to similar selection pressures in both I factors and L1 elements.

Introns found in mitochondrial genes have been classified as being of class I or class II according to their sequence, and potential secondary structures (Michel and Dujon, 1983). Four class II introns contain open reading frames that can be aligned along their length and have some homology with the conserved regions in reverse transcriptases (Michel and Long, 1985). Two of these, the

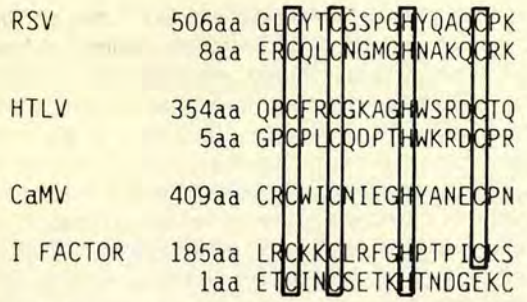


Figure 7. Comparison of Part of ORF1 of the I Factor with Conserved Domains in Viral Nucleic Acid Binding Proteins

The viral sequences are as follows: RSV, the *gag* gene of RSV (Schwartz et al., 1983); HTLV, the *gag* gene of HTLVI (Seiki et al., 1983); CaMV, the coat protein of CaMV (Franck et al., 1980). The I factor sequence is from ORF1. The numbers indicate the number of amino acid residues from the start of the polypeptide or the number of residues between the two blocks of sequence in a single polypeptide. Conserved residues are boxed.

sequences encoded by the a1 and a2 introns of the S. cerevisiae cytochrome oxidase subunit I gene, are shown in Figure 6. In region 2, where the I factor and L1 elements are most closely related, the class II intron sequences are more like those of the I factor and L1 elements than the viral reverse transcriptases.

Two families of elements with some properties similar to those of the I factor and L1 elements have already been found in D. melanogaster. These are known as F and G elements (Dawid et al., 1981; Pardue and Dawid, 1981; Di Nocera and Dawid, 1983). There are only 10–20 G elements per haploid genome, many of which are in tandem arrays inserted in the nontranscribed spacer of rDNA units (Di Nocera and Dawid, 1983; Di Nocera et al., 1986). They have no terminal repeats, but do have an A-rich sequence at the 3′ end of one strand. The chromosomal distribution of G elements is fairly stable, and they are concentrated in chromocentric regions.

The family of F elements has approximately 50 members located at the chromocenter and about 25 sites on chromosome arms (Dawid et al., 1981). They appear to be transposable since the distribution of euchromatic sites is different in different strains and individual elements are flanked by target site duplications differing in length from element to element (Di Nocera et al., 1983). F elements, like I factors, have no terminal repeats and have an A-rich sequence at the 3′ end of one strand. They vary in length, many copies being deleted at the 5′ end; however, a consensus restriction map 4.7 kb long has been constructed. About 800 bp from the ends of an apparently full-length F element, 101F, have been reported (Di Nocera et al., 1983), but despite the structural similarity between these elements and the I factor, we can detect homology neither between this DNA sequence and the I factor, nor between the open reading frames in 101F and either ORF1 or ORF2. The complete sequence of a full-length F element may well reveal an open reading frame related to reverse transcriptases, like those described above for the I factor and L1 elements.

The polypeptides encoded by retroviral *pol* genes are not only associated with reverse transcriptase activity, they also contain a region that serves as an endonuclease required for proviral integration and, in some cases, a segment that acts as a protease, cleaving viral polypeptides into their components. We can find no regions in ORF1 or ORF2 that are similar to these endonuclease or protease domains.

We have compared ORF1 with version 5 of the National Biomedical Research Foundation Protein Data Base, as described in Experimental Procedures, to try to detect similar sequences. No proteins showed striking homology to ORF1, but H. Pinon has drawn our attention to the fact that within ORF1 there is a sequence that precisely matches a highly conserved motif, $CX_2CX_4HX_4C$, found in basic nucleic acid–binding proteins cleaved from retroviral *gag* polypeptides. This is thought to interact directly with genomic RNA (Dickson et al., 1985; Copeland et al., 1984; Covey, 1986). Covey has searched for this sequence within polypeptides encoded by a number of elements known, or believed, to use a reverse transcriptase. It is present in all retroviral *gag* polypeptides tested, in polypeptides coded by the corresponding regions of intracisternal A-type particle DNA of Syrian hamster and the *copia* elements of D. melanogaster, and in the viral coat protein of CaMV, but not in the *copia*-like element *17.6*, the Ty element of S. cerevisiae, or human HBV. Some genes contain two copies of this sequence; others only one. ORF1 has one complete and one partial copy. These are shown in Figure 7 aligned with the conserved domains from the *gag* genes of RSV, HTLVI, and CaMV. The presence of a potential nucleic acid–binding domain in ORF1 suggests that the corresponding polypeptide may interact with I factor DNA or RNA, possibly playing a role in I factor regulation or transposition.

Discussion

We have determined the complete base sequence of the I factor associated with the w^{IR1} mutation and have analyzed both ends of six other I factors. Each element is flanked by a target site duplication. These vary slightly in length, and we have found duplications of 10–14 bp so far. Insertion of I factors seems to have some sequence specificity since three elements have inserted at exactly the same position (Figure 3). This specificity could be due to a preferred target site sequence or to some feature of the adjacent DNA. There is no striking homology between the target sites we have identified so far.

The structure of the I factor differs from that of most other transposable elements in that it has no direct or inverted terminal repeats. It is quite different from the P factor, in particular, confirming that the P–M and I–R systems of hybrid dysgenesis are quite distinct. The 3′ end of one strand of the I factor is A-rich, having 4–7 tandem repeats of the sequence TAA, and this strand contains two long open reading frames, ORF1 and ORF2. This sequence organization is similar to that of LINEs, or L1 elements, in mammals, and probably that of F elements in D. melanogaster. The A-rich 3′ sequences of L1 and F elements are thought to be created by polyadenylation of full-length

RNAs since they are preceded, at a suitable distance, by the polyadenylation signal AATAAA. This is not true of I factors. The TAA repeats are not likely to be poly(A) sequences that have been degraded by mutation, since the pattern is the same in each of the seven I factors we have analyzed and the nearest polyadenylation signal is 73 bp upstream. This would be suitable for polyadenylation of a messenger RNA encoding ORF2, but not of full-length RNA. We think that the TAA repeats may be the end of transposition intermediates that integrate at staggered nicks in chromosomal DNA and that they may be increased or decreased in number by slippage during the polymerization and ligation that takes place to generate a target site duplication. Another possibility is that they change in number as a result of unequal recombination.

L1 elements are believed to transpose by reverse transcription of a polyadenylated RNA to give an extrachromosomal DNA intermediate that integrates at a new site in the genome. The fact that some L1 elements are known to encode a polypeptide related to viral reverse transcriptases is consistent with this. I factors are structurally related to L1 elements, are known to determine a function required for their own transposition, and encode a polypeptide with some similarity to reverse transcriptases. For these reasons, we think that I factors also transpose via an RNA intermediate. This raises two questions. How is this RNA transcribed, and what is the primer for reverse transcription?

We have determined the sequences at both ends of each of seven I factors and have found them to be highly conserved. If transposition does involve an RNA intermediate, then this RNA must be able to include the complete sequence of an I factor. The promoter for synthesis of this RNA must be carried within the I factor since we know that the element associated with the w^{IR1} mutation can transpose to new sites (Pelisson, 1981). It cannot use the *white* gene promoter since the I factor would have to be transcribed in the opposite direction. The promoter must therefore be carried within the I factor itself. The I factor is presumably transcribed by RNA polymerase II, since we can find no sequences related to box A and box B of pol III promoters (Galli et al., 1981) in the first 200 bp of the I factor and both strands contain several oligo(T) regions known to serve as pol III terminators (Bogenhagen and Brown, 1981). We are forced to conclude that if the I factor transposes using an RNA intermediate, then this must be transcribed from an internal pol II promoter. The mRNAs for ORF1 and ORF2 could be generated by processing full-length RNA transcribed from this promoter or transcribed from their own promoters.

The position of the promoters for transposition intermediates of L1 elements is less certain since one cannot determine whether any particular element can transpose. There could be a "master" L1 element, or elements, in the genome that has an external promoter and that is the source of full-length RNAs. Loeb et al. (1986) have found that the 5′ (left-hand) end of the mouse L1 element L1Md-A2, has 4⅔ tandem repeats of a 208 bp sequence containing regions with some similarity to the promoters of several housekeeping genes and of SV40. They suggest that these repeats contain the promoter for L1Md and that transcription starts about 70 bp from the beginning of a repeat. There are no such repeats at the left-hand end of the sequence of the I factor or the sequences reported for long human and rat L1 elements (Hattori et al., 1986; D'Ambrosio et al., 1986). Skowronski and Singer (1985) have detected full-length L1 RNAs in human teratocarcinoma cells, raising the possibility that human L1 elements may also have an internal pol II promoter.

We have tried to identify possible primers for reverse transcription, but so far without success. We have been unable to detect any suitable homology between the 3′ ends of any of the D. melanogaster tRNAs and other small RNAs for which sequence information is available (GENBANK release 40; Sprinzl et al., 1985). We have also investigated the possibility that the 3′ end of a full-length RNA transcribed from left to right in Figure 1, might fold back to form a loop, which could act as a primer. Some potentially stable secondary structures can be formed within the last 200 bases, but none of these seem to be appropriate for priming DNA synthesis.

The fact that retrotransposons (Boeke et al., 1985), L1 elements, and I factors may all transpose using reverse transcriptases encoded by these elements themselves does not necessarily indicate that they are related by descent. Each may have evolved separately from nontransposable reverse transcriptase genes. Such genes would have the potential to replicate independently of other chromosomal sequences since reverse transcriptase could convert its own mRNA into DNA, the two being in close proximity immediately after translation. This would only require a mechanism for priming DNA synthesis, and gene copies generated in this way could enter the nucleus and integrate into the genome. These would be chance events initially, and might still be so for elements that transpose infrequently. Elements that transpose at high frequency under appropriate circumstances, like the I factor, must have evolved efficient mechanisms for both processes.

Experimental Procedures

Bacteria, Bacteriophages, and Drosophila Strains

Recombinant lambda NM762 phages (Murray et al., 1977; Williams and Blattner, 1980) were plated on E. coli strain ED8654 (*hsd*R, *met*B, *sup*E, *sup*F) (Borck et al., 1976), and recombinant lambda NM1149 phages (Murray, 1983a) were plated on NM514 (*hsd*R, *lyc*7) (Murray, 1983b). The host for M13 phages was NM522 (Δ*lac-pro*, *hsd*MS, F′*lac*ZM15,*lacI*^{q}) (Gough and Murray, 1983).

All Drosophila strains were from the Laboratoire de Genetique, Universite de Clermont-Ferrand.

Enzymes and Isotopes

Restriction enzymes were purchased from Boehringer Mannheim, Pharmacia, and Amersham International; E. coli DNA polymerase I was from Boehringer Mannheim; Klenow fragment was from Boehringer Mannheim and Pharmacia; T4 DNA ligase was from New England Biolabs, [α-^{32}P]dCTP (3,000 Ci/mMol) was from Amersham International; and [α-^{35}S]dATP was from New England Nuclear (500 Ci/mMol) and Amersham International (410 Ci/mMol).

DNA Preparation, In Vitro Labeling, Hybridization, and Autoradiography

Phage DNA was prepared from liquid lysates as described by Will et

al. (1981). D. melanogaster DNA was prepared as described by Bucheton et al. (1984).

In vitro labeling of DNA, hybridization, and autoradiography were carried out as described by Will et al. (1981). Hybridization filters were washed for 1 hr in 2× SSC and 0.1% SDS at 37°C, then for 1 hr in 2× SSC at room temperature.

Construction of Libraries

A library of cloned HindIII fragments of w^{IR4} DNA was constructed by ligating 1 μg of HindIII-cut lambda NM1149 DNA with 1.5 μg of HindIII-cut w^{IR4} DNA. Ligation was carried out overnight at 10°C in 10 μl of 66 mM Tris–HCl (pH 7.2), 1 mM EDTA, 10 mM $MgCl_2$, 10 mM dithiothreitol, 0.1 mM ATP, and 100 units T4 DNA ligase. Libraries of cloned HindIII fragments of w^{IR6} and w^{IR8} DNAs were constructed in a similar way.

A library of w^{IR7} DNA was constructed in lambda NM1149 as described for w^{IR4}, except that EcoRI DNA was used.

A library of w^{IR6} HindIII fragments was constructed in the replacement vector lambda NM762 by ligating 1 μg of HindIII-cut NM762 DNA with 1.5 μg of w^{IR6} DNA partially digested with HindIII, using the conditions described above. This was used to clone the right-hand end of this I factor.

Following ligation, the libraries were packaged in vitro as described by Scherer et al. (1981) and screened with appropriate fragments of DNA from the *white* locus.

DNA Sequencing

The data in Figure 2 were obtained as follows. Plasmids carrying the DNA to be sequenced were sonicated (Deininger, 1983), and the ends of the fragments produced were repaired using DNA polymerase I. Fragments of 200–1000 bp were isolated by gel electrophoresis and cloned into the SmaI site of M13mp9 or M13mp10. Templates were made and sequenced by the dideoxynucleotide chain termination method (Sanger et al., 1977) using [^{35}S]dATP and were analyzed on buffer gradient polyacrylamide gels (Biggin et al., 1983). The sequences from the ends of I factors and from the $w^{IR7,8}$ deletions were obtained by cloning appropriate restriction fragments into M13 vectors. In some cases, oligonucleotide primers were synthesized to sequence-specific regions.

Sequence Analysis

Random sequencing data were assembled using the DataBase programs of Staden (1982). DNA and protein sequences were analyzed using programs written by Devereux et al. (1984). The sequences of ORF1 and ORF2 of the I factor were compared with the NBRF data base by J. Collins and A. Lyall using the "Prelate" system (Collins and Coulson, 1986).

Acknowledgments

We are grateful to G. Brown for photography, S. Walsh for technical assistance, C. Leaver for oligonucleotides, K. O'Hare for material relating to the w^1 mutation, M. Peifer and W. Bender for the bx^{F31} clone, J. Collins and A. Lyall for computer analysis, A. Coulson for computer assistance, H. Pinon and M. Singer for stimulating discussions, F. Burton for drawing our attention to CS1, and A. Bucheton, E. Livingstone, A. Pelisson, and H. Pinon for comments on the manuscript.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 10, 1986; revised October 10, 1986.

References

Biggin, M. D., Gibson, T. J., and Hong, G. F. (1983). Buffer gradient gels and ^{35}S label as an aid to rapid sequence determination. Proc. Natl. Acad. Sci. USA *80*, 3963–3965.

Boeke, J. D., Garfinkel, D. J., Styles, C. A., and Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491–500.

Bogenhagen, D. F., and Brown, D. D. (1981). Nucleotide sequences in Xenopus 5S DNA required for transcription termination. Cell *24*, 261–270.

Bonitz, S. G., Coruzzi, G., Thalenfeld, B. E., Tzagoloff, A., and Macino, G. (1980). Assembly of the mitochondrial membrane system, structure and nucleotide sequence of the gene coding for subunit I of cytochrome oxidase. J. Biol. Chem. *255*, 11927–11941.

Borck, K., Beggs, J. D., Brammar, W. J., Hopkins, A. S., and Murray, N. E. (1976). The construction of in vitro transducing derivatives of phage lambda. Mol. Gen. Genet. *196*, 199–207.

Bregliano, J. C., and Kidwell, M. G. (1983). Hybrid dysgenesis determinants. In Mobile Genetic Elements, J. Shapiro, ed. (New York: Academic Press), pp. 363–410.

Bucheton, A., Paro, R., Sang, H. M., Pelisson, A., and Finnegan, D. J. (1984). The molecular basis of I–R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. Cell *38*, 153–163.

Burton, F. H., Loeb, D. D., Voliva, C. F., Martin, S. L., Edgell, M. H., and Hutchison, C. A. (1986). Conservation throughout mammalia and extensive protein encoding capacity of the highly repeated DNA long interspersed sequence one. J. Mol. Biol. *187*, 291–304.

Chiu, I. M., Yaniv, A., Dahlberg, J. E., Grazit, A., Skuntz, S. F., Tronick, S. R., and Aaronson, S. A. (1985). Nucleotide sequence evidence for relationship of AIDS retrovirus to lentiviruses. Nature *317*, 366–368.

Clare, J., and Farabaugh, P. (1985). Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. Proc. Natl. Acad. Sci. USA *82*, 2829–2833.

Collins, J. F., and Coulson, A. F. W. (1986). Molecular sequence comparison and alignment. In Nucleic Acid and Protein Sequence Analysis: A Practical Approach, M. Bishop and C. Rawlings, eds. (Oxford: IRL Press), in press.

Copeland, T. D., Orozlan, S., Kalyanaramam, V. S., Sarngadharan, M. G., and Gallo, R. C. (1983). Complete amino acid sequence of human T cell leukemia virus structural protein p15. FEBS Lett. *162*, 390–395.

Covey, S. N. (1986). Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucl. Acids Res. *14*, 623–633.

D'Ambrosio, E., Waitzkin, S. D., Witney, F. R., Salemme, A., and Furano, A. V. (1986). Structure of highly repeated, long interspersed DNA family (LINE or L1Rn) of the rat. Mol. Cell. Biol. *6*, 411–424.

Dawid, I. B., Long, E. O., Di Nocera, P. P., and Pardue, M. L. (1981). Ribosomal insertion-like elements in Drosophila melanogaster are interspersed with mobile sequences. Cell *25*, 399–408.

Deininger, P. I. (1983). Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal. Biochem. *129*, 215–223.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. Nucl. Acids Res. *12*, 387–395.

Dickson, C., Eisenman, R., and Fan, H. (1985). Protein biosynthesis and assembly. In RNA Tumor Viruses, R. Weiss, N. Teich, H. Varmus, and J. Coffin, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory), pp. 135–145.

Di Nocera, P. P., and Dawid, I. B. (1983). Interdigitated arrangement of two oligo(A)-terminated DNA sequences in *Drosophila*. Nucl. Acids Res. *11*, 5475–5482.

Di Nocera, P. P., Digan, M. E., and Dawid, I. (1983). A family of oligoadenylated transposable sequences in *Drosophila melanogaster*. J. Mol. Biol. *168*, 715–727.

Di Nocera, P. P., Graziani, F., and Lavorgna, G. (1986). Genomic and structural organisation of *Drosophila melanogaster* G elements. Nucl. Acids Res. *14*, 675–691.

Emori, Y., Shiba, T., Kanaya, S., Inouye, S., Yuki, S., and Saigo, K. (1985). The nucleotide sequences of *copia* and *copia*-related RNA in *Drosophila* virus-like particles. Nature *315*, 773–776.

Finnegan, D. J. (1985). Transposable elements in eukaryotes. Int. Rev. Cytol. *93*, 281–326.

Franck, A., Guilley, H., Jonard, G., Richards, K., and Hirth, L. (1980). Nucleotide sequence of cauliflower mosaic virus DNA. Cell *21*, 285–294.

Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P., and Charnay, P. (1979). Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. Nature *281*, 646–650.

Galli, G., Hofstetter, H., and Birnstiel, M. L. (1981). Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. Nature *294*,, 626–631.

Gardner, R. C., Howarth, A. J., Hahn, P., Brown-Luedi, M., Shepherd, R. J., and Messing, J. (1981). The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucl. Acids Res. *9*, 2871–2887.

Gebhardt, W., and Zachau, H. (1983). Organization of the R family and other interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. *170*, 255–270.

Gough, J. A., and Murray, N. E. (1983). Sequence diversity among related genes for recognition of specific targets in DNA molecules. J. Mol. Biol. *166*, 1–19.

Hattori, M., Kuhara, S., Takenaka, O., and Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase–related protein. Nature *321*, 625–628.

Jagadeeswaran, P., Forget, B. G., and Weissman, S. M. (1981). Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA pol III transcripts? Cell *26*, 141–142.

Kidwell, M. G. (1983). Intraspecific hybrid sterility. In Genetics and Biology of Drosophila, Vol. 3c, M. Ashburner, H. L. Carson, and J. N. Thompson, eds. (London: Academic Press), pp. 125–153.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucl. Acids Res. *12*, 857–872.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiation codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283–292.

Loeb, D. D., Padgett, R. W., Hardies, S. C., Shehee, W. R., Comer, M. B., Edgell, M. H., and Hutchison, C. A. (1986). The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retrotransposons. Mol. Cell. Biol. *6*, 168–182.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eucaryotic DNA. Cell *15*, 687–701.

Michel, F., and Dujon, B. (1983). Conservation of RNA secondary structure in two intron families including mitochondrial, chloroplast and nuclear encoded members. EMBO J. *2*, 33–38.

Michel, F., and Lang, B. F. (1985). Mitochondrial class two introns encode proteins related to the reverse transcriptases of retroviruses. Nature *316*, 641–643.

Mount, S. M., and Rubin, G. M. (1985). Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. Mol. Cell. Biol. *5*, 1630–1638.

Murray, N. E. (1983a). Lambda vectors. In Lambda II, R. W. Hendrix, J. W. Roberts, F. W. Stahl, and R. A. Weisberg, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory), pp. 677–684.

Murray, N. E. (1983b). Phage lambda and molecular cloning. In Lambda II, R. W. Hendrix, J. W. Roberts, F. W. Stahl, and R. A. Weisberg, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory), pp. 395–432.

Murray, N. E., Brammar, W. J., and Murray, K. (1977). Lambdoid phages that simplify the recovery of *in vitro* recombinants. Mol. Gen. Genet. *150*, 53–61.

O'Hare, K., and Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome. Cell *34*, 25–35.

O'Hare, K., Levis, R., and Rubin, G. M. (1983). Transcription of the *white* locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *80*, 6917–6921.

O'Hare, K., Murphy, C., Levis, R., and Rubin, G. M. (1984). DNA sequence of the *white* locus of *Drosophila melanogaster*. J. Mol. Biol. *180*, 437–455.

Pardue, M. L., and Dawid, I. B. (1981). Chromosomal locations of two DNA segments that flank ribosomal insertion-like sequences in *Drosophila*: flanking sequences are mobile elements. Chromosoma *83*, 29–43.

Patarca, R., and Heseltine, W. A. (1984). Letter. Nature *309*, 288.

Peifer, M., and Bender, W. (1986). The anterio-bithorax and bithorax mutations of the bithorax complex. The EMBO J., in press.

Pelisson, A. (1981). The I–R system of hybrid dsygenesis in *Drosophila melanogaster*: are I factor insertions responsible for the mutator effect of the I–R interaction? Mol. Gen. Genet. *183*, 123–129.

Rogers, J. E. (1983). Retroposons defined. Nature *302*, 460.

Rogers, J. E. (1985). The origin and evolution of retroposons. Int. Rev. Cytol. *93*, 188–280.

Saigo, K., Kugimya, W., Matsuo, Y., Inouye, S., Yoshioka, K., and Yuki, S. (1984). Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature *312*, 659–661.

Sang, H. M., Pelisson, A., Bucheton, A., and Finnegan, D. J. (1984). Molecular lesions associated with *white* gene mutations induced by I–R hybrid dysgenesis in *Drosophila melanogaster*. EMBO J. *3*, 3079–3085.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.

Scherer, G., Telford, J., Baldari, C., and Pirrotta, V. (1981). Isolation of cloned genes differentially expressed at early and late stages of *Drosophila* embryonic development. Dev. Biol. *86*, 438–447.

Schwartz, D. E., Tizard, R., and Gilbert, W. (1983). Nucleotide sequence of Rous sarcoma virus. Cell *32*, 853–869.

Seiki, M., Hattori, S., Hiroyama, Y., and Yoshida, M. (1983). Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sci. USA *80*, 3618–3622.

Singer, M. F. (1982). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell *28*, 433–434.

Singer, M. F., and Skowronski, J. (1985). Making sense out of LINEs: long interspersed repeat sequences in mammalian genomes. Trends Biochem. Sci. *10*, 119–122.

Skowronski, J., and Singer, M. F. (1985). Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc. Natl. Acad. Sci. USA *82*, 6050–6054.

Sprinzl, M., Vorderwulbecke, T., and Hartmann, T. (1985). Compilation of tRNA sequences. Nucl. Acids Res. *13* suppl., r51-r104.

Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucl. Acids Res. *10*, 4731–4751.

Toh, H., Hayashida, H., and Miyata, T. (1983). Sequence homology between retroviral reverse transcriptase and putative polymerase of hepatitis B virus and cauliflower mosaic virus. Nature *305*, 827–829.

Toh, H., Kikuno, R., Hayashida, T., Kugimiya, W., Inouye, S., Yuki, S., and Saigo, K. (1985). Close structural resemblance between putative polymerase of a *Drosophila* transposable genetic element *17.6* and *pol* gene product of Moloney murine leukemia virus. EMBO J. *4*, 1267–1272.

Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T., and Gesteland, R. F. (1981). Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell *26*, 11–17.

Will, B. M., Bayev, A. A., and Finnegan, D. J. (1981). Nucleotide sequence of terminal repeats of 412 transposable elements of *Drosophila melanogaster*. J. Mol. Biol. *153*, 897–915.

Williams, B. G., and Blattner, F. R. (1980). Bacteriophage lambda vectors for DNA cloning. In Genetic Engineering, Vol. 2, J. K. Setlow and A. Mullander, eds. (New York: Plenum Press), pp. 201–281.

Zachar, Z., and Bingham, P. M. (1982). Regulation of *white* locus expression: the structure of mutant alleles at the *white* locus of Drosophila melanogaster. Cell *30*, 529–541.

An Offprint from

OXFORD SURVEYS ON
EUKARYOTIC GENES

Volume 3 : 1986

1 Transposable elements in *Drosophila melanogaster*

DAVID J. FINNEGAN AND
DIANA H. FAWCETT

I. Introduction

Transposable elements have been detected in the genomes of species representing virtually all forms of life. Amongst eukaryotes, by far the largest number and greatest variety have been found in the genome of *Drosophila melanogaster* (Table 1.1). They are all moderately repetitive sequences and make up about two-thirds of this fraction of the genome, that is, at least 10 per cent of the mass of the DNA (Young 1979). *D. melanogaster* is probably unusual only in the proportion of its genome comprising transposable sequences, not in their variety, and it will be surprising if any of the transposable elements found in this species do not have counterparts elsewhere.

These elements can be classified according to their sequence organization, and, in general, this is how they are discussed in this review. We have tried to organize this review so that it will be useful to both the general reader and those seeking specific detailed information. In the text we discuss the properties of each class of transposable element and briefly consider some of the general problems raised by them. The appendix to this chapter contains detailed information about those elements which have been characterized. The book *Mobile Genetic Elements*, edited by Shapiro (1983), deals with transposable elements in general, and the reviews by Georgiev (1984) and Finnegan (1985) discuss transposable elements in eukaryotes.

II. *Copia*-like elements—transposable elements with long terminal direct repeats

A. THE PROPERTIES OF *COPIA*-LIKE ELEMENTS

The first transposable DNA sequences to be detected in *D. melanogaster*, or indeed any eukaryote, were members of the class of elements now referred to as *copia*-like elements (Rubin *et al.* 1976; Georgiev *et al.* 1977). These occur as families of repeated sequences, each family being named after the first member to be studied (see appendix). Each family has about 10–50 members per haploid genome, depending on the strain of flies in question. The number is generally greater in tissue culture cells where there are often over 100 copies per haploid genome (Potter *et al.* 1979; Ilyin *et al.* 1980a, b). The best-characterized family is the *copia* family, and it is for this reason that similar

Table 1.1

Transposable elements of D. melanogaster ordered according to size. More information about these elements can be found in the text and/or the appendix.

Element	Length (kb)
Sancho 2	2.6
Jockey	2.8
P	2.9
hobo	3.0
Doc	4.3
1731	4.4
Kermit	4.8
Sancho 1	4.5
copia	5.1
I	5.4
mdg3	5.4
NEB	5.5
3S18	6.5
297	7.0
Delta 88	7.0
Calypso	7.2
Harvey	7.2
BEL	7.3
HMS Beagle	7.3
mdg1	7.3
mdg4/gypsy	7.3
17.6	7.4
412	7.6
BS	8.0
B104/roo	8.7
springer	8.8
F	variable
FB	variable
G	variable

elements are usually called '*copia*-like' (Rubin *et al.* 1981), although some authors refer to them as *m*obile *d*isperse *g*enes, or *mdg* elements (Bayev *et al.* 1980).

The most characteristic feature of *copia*-like elements is that they have long direct repeat sequences at their termini (Finnegan *et al.* 1978). These are known as 'long terminal repeats' or 'LTRs'. The length of these repeats is constant within a family and varies from about 250 to 500 base pairs (bp) between families. The base sequences of the LTRs of several families have been determined and are given in the appendix. These sequences are well conserved, and are usually identical at the ends of any particular copy of an element (Levis *et al.* 1980; Will *et al.* 1981). There is no extensive sequence

homology between the LTRs of different families with the exception of the *17.6* and *297* families, which appear to be evolutionarily related (Kugimya *et al.* 1983).

There is no DNA sequence homology between members of different families, except for the *17.6* and *297* families, but elements are very similar to each other within a family. This can be demonstrated by using an internal restriction fragment of an element to probe an appropriate digest of genomic DNA in a Southern transfer experiment. A strong band of hybridization is seen corresponding to the internal fragment of the element. The intensity of hybridization of this band gives a measure of the number of copies present. This is illustrated for the *412* family in Fig. 1.1a.

The first suggestion that *copia*-like elements might be transposable came from *in situ* hybridization experiments illustrating the chromosomal distribution of members of particular families. In each case there was hybridization to sites on the chromosome arms and centromeric heterochromatin, indicating that the elements were repeated and scattered throughout the genome (Rubin *et al.* 1976; Georgiev *et al.* 1977; Finnegan *et al.* 1978; Ilyin *et al.* 1978). When these sites of hybridization were compared between strains, they were found to be very different (Ilyin *et al.* 1978; Potter *et al.* 1979). This can only easily be explained if the elements in question are able to change their chromosome location, that is, are transposable.

This conclusion is supported by the results of Southern transfer experiments in which digests of genomic DNA from different strains of *D. melanogaster* are probed so as to detect restriction fragments including the ends of each member of a family (Strobel *et al.* 1979). The size of end fragments will be determined by the position of restriction sites in the adjacent DNA and will usually differ from one copy of an element to another. The pattern of end fragments in DNA from any strain will depend on the chromosomal distribution of elements. When the patterns from several strains are compared they are found to be very different, reflecting differences in the distribution of elements from strain to strain. This is illustrated in Fig. 1.1b for members of the *412* family. Different hybridization patterns could, in principle, be produced by DNA from strains with identical distributions of elements, but with very heterogeneous restriction sites in adjacent DNA. This may be responsible for some of the differences seen, but is not generally the case.

The only direct way to demonstrate that sequences are transposable is to observe their movement within the genome of a particular strain. This can be recognized most easily when elements transpose into genes causing mutations. This was first demonstrated by Bingham and Judd (1981). They used a sophisticated combination of molecular and genetic techniques to show that a *copia* element is inserted within the *white* gene on chromosomes carrying the *white*-apricot mutation, w^a. There are now many such examples (see appendix) and it appears that about one-half of spontaneous mutations in *D. melanogaster* are associated with insertion of *copia*-like elements (Zachar and

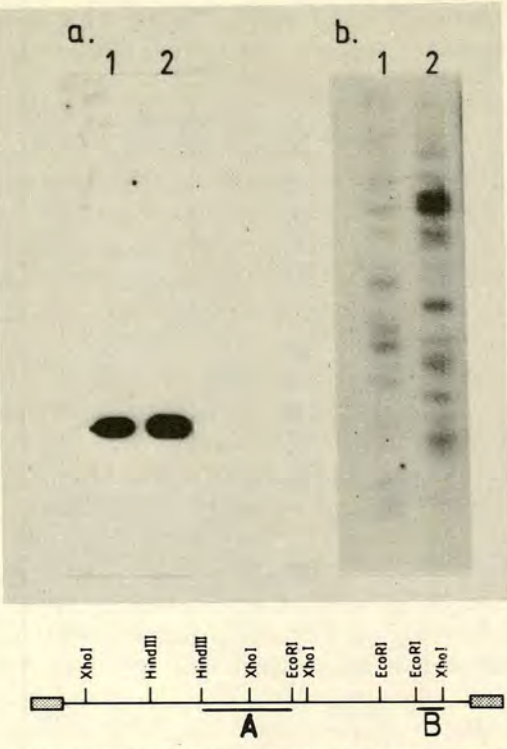


Fig. 1.1 Restriction digests of *Drosophila* genomic DNA probed with fragments of the *copia*-like element *412*. (a) Genomic DNAs of the strains w^{IR1}, 1, and *iso5*, 2, were digested with both HindIII and EcoRI, fractionated on a 1 per cent agarose gel, and then transferred to nitrocellulose. The DNAs were then hybridized with a sub-clone of the internal HindIII.EcoRI fragment, A, of a *412* element. Note that there is only a single band of hybridization. This corresponds to fragments of the same size as fragment A, indicating that *412* elements are highly conserved in this region. (b) Genomic DNAs of the same strains as above were digested with EcoRI and fractionated on a 0.7 per cent agarose gel. They were then transferred to nitrocellulose and hybridized with a sub-clone of the EcoRI.XhoI fragment, B, of a *412* element. Note that many fragments hybridize to this probe in the DNA of both strains, but that the two patterns are different. This indicates that the chromosomal distribution of *412* elements is different in these strains. These experiments were carried out as described by Bucheton *et al.* (1984). The restriction map of a *412* element at the bottom of the figure shows the two fragments used as probes.

Bingham 1982; Bender *et al.* 1983*a*; Mattox and Davidson 1984: O'Hare *et al.* 1984).

One of the most characteristic features of transposable elements is that they are flanked by direct repeats of a small number of bases. These bases occur only once at the target site for insertion prior to arrival of an element, and they are believed to be duplicated as a result of the mechanism of transposition. The number of bases duplicated is usually characteristic of a family

of elements, but their sequence may vary from one copy of the element to another. This is true of *copia*-like elements which transpose more or less randomly, although *17.6* and *297* elements seem to insert preferentially at the sequence ATAT (Rubin 1983; Inouye *et al.* 1984) and possible hot-spots have been found for *copia* insertion in the *white* gene (Rubin *et al.* 1982), and for *gypsy* insertion in the *scute* gene (Campuzano *et al.* 1985). Some chromosome regions, particularly those which are heterochromatic, appear to have a high concentration of *copia*-like elements (Ananiev *et al.* 1978; Tchurikov *et al.* 1980; Young and Schwartz 1981; Montgomery and Langley 1983). These could be preferential targets for insertion, either because of their DNA sequence or because of some feature of chromatin structure. Alternatively, they may be regions of the genome which can accumulate transposable elements without deleterious effect on the organism as a whole.

The very heterogeneous distribution of *copia*-like elements in different strains of *D. melanogaster* indicates that they are lost by excision. The rate of excision must be about the same as that of insertion since the number of copies of any particular element is about the same in different strains (Young 1979), but the two processes need not be related mechanistically. Recombination between the LTRs at the ends of an element will excise a circular molecule containing one LTR and leave the second 'free' or 'solo' in the chromosome. Free LTRs do occur, and free *412* (Shepherd and Finnegan 1984) and *mdg*3 (Mossie *et al.* 1985) LTRs have been cloned, but they are rare (Levis *et al.* 1980; Kulguskin *et al.* 1981; Shepherd and Finnegan 1984). Presumably recombination between LTRs is a rare event, or it is followed soon afterwards by loss of the resulting free LTR. Mutations bx^3 and bxd^1 are associated with insertion of members of the *gypsy* family of *copia*-like elements (Bender *et al.* 1983*a*). Members of this family had been described previously by Tchurikov *et al.* (1978) who called them *mdg*4 elements (Bayev *et al.* 1984). Phenotypic revertants of these mutations have been found which have lost the bulk of the *gypsy* elements, retaining only a single LTR. This is presumably the result of LTR–LTR recombination. A similar phenomenon has been found for mutations associated with *gypsy* insertions in the *scute* gene (Campuzano *et al.* 1985).

Mutations associated with insertion of *copia*-like elements are generally stable, indicating that precise excision must be a rare event. This is not necessarily true of LTR–LTR recombination, since the effect of a free LTR on gene expression will depend very much on its position with respect to the gene in question. Insertions within introns might be expected to give phenotypic revertants more frequently than those within exons, although the effects of DNA insertions in non-coding regions are difficult to predict. Carbonara and Gehring (1985) have studied four derivatives of the w^a mutation. One of these, w^{aR59K1}, retains a *copia* LTR at the original site of insertion, presumably the result of LTR–LTR recombination. This rearrangement has only partially reversed the phenotypic effect of w^a, even

though the original *copia* insertion was within an intron. The *copia* insertion itself affects the size of transcripts from the *white* gene (Levis *et al.* 1984; Pirotta and Brock 1984), probably because there is a polyadenylation signal, AATAAA, within the LTRs (see below). This should be present in the single LTR in w^{aR59K1}, and may be responsible for its slight mutant effect.

Most *copia*-like elements are transcribed into polyadenylated RNAs found in the cytoplasm of tissue culture cells and various developmental stages of the fly. These RNAs can represent up to 3 per cent of the cytoplasmic polyA$^+$ RNA (Finnegan *et al.* 1978; Ilyin *et al.* 1978; Falkenthal and Lengyel 1980), and many *copia*-like elements were first identified as abundantly transcribed genes. At least some of these RNAs are about the full length of the element from which they were transcribed, and in the case of *copia* (Flavell *et al.* 1980) and *B104* (Scherer *et al.* 1982) these have been shown to have their 5′ and 3′ ends within the corresponding LTRs. Several smaller *copia* RNAs have been detected including a prominent 2 kilobase (kb) species (Flavell *et al.* 1980; Schwartz *et al.* 1982). A number of RNAs have been found for *mdg*1 (Ilyin *et al.* 1980*c*), *mdg*3 (Ilyin *et al.* 1980*a*), *412* (Schwartz *et al.* 1982) and *B104* (Scherer *et al.* 1982) elements.

The transcripts from *copia*-like elements are not found on polysomes to any significant extent (Flavell *et al.* 1980; Ilyin *et al.* 1980c; Falkenthal and Lengyel 1980; Young and Schwartz 1981). *Copia* RNAs are translated inefficiently by heterologous *in vitro* systems (Flavell *et al.* 1980; Shiba and Saigo 1983) to give a heterogenous population of polypeptides. The largest product of *in vitro* translation reported by Flavell *et al.* (1980) was 51 000 atomic mass units (u) and was stimulated by the 2 kb RNA. Shiba and Saigo (1983) have found a prominent 31 000 u product using *copia* RNA prepared in a different way.

The level of transcription of several *copia*-like elements has been found to vary during development. Schwartz *et al.* (1982) measured the level of *copia* and *412* transcripts in total cellular RNA from embryos, larvae and adults of three different wild-type strains of *D. melanogaster*. The levels of both *copia* and *412* RNAs varied during development but was not co-ordinate. *Copia* expression was highest in adults and larvae, whereas *412* RNA was most abundant in embryos. These patterns were the same for each of three different wild-type strains. This result suggests that *copia*-like elements regulate their transcription autonomously, rather than being controlled by adjacent sequences, since the chromosomal distribution of *copia* and *412* elements was probably very different in each of these strains. The picture is not entirely clear since Flavell *et al.* (1980), using another strain, could detect *copia* RNA in larvae but not embryos or adults. Most *copia*-like elements are expressed at high levels in tissue culture cells.

The idea that expression of *copia*-like elements is controlled by the elements themselves, rather than by surrounding chromatin, is supported by the observations of Meyerowitz and Hogness (1982) concerning an element near

an abundantly transcribed gene. They found a strain of *D. melanogaster* in which a member of the *B104/roo* family of elements is present close to the glue protein gene *sgs*3. This gene is expressed at high levels in salivary glands of late third instar larvae. Meyerowitz and Hogness could detect *roo* transcripts in embryos of this strain but not in RNA isolated from larval salivary glands, even when *sgs*3 was expressed and the chromosomal region containing both sequences was puffed. Either this particular *roo* element is incapable of expression, or it is regulated independently of *sgs*3.

Studies of the expression of *copia*-like elements are complicated by the fact that there is as yet no information to indicate how many members of any particular family are expressed at any one time. One way of investigating this question would be to transform flies with a marked element and to follow its expression during development. The effect of chromosome position on expression could be studied by assaying transformants with the marked element integrated at different sites.

Expression of some *copia*-like elements appears to be modulated by *trans*-acting products coded by genes elsewhere in the genome. The recessive mutation *su(Hw)* (Lewis 1949) can suppress the mutant effects of alleles of genes scattered throughout the genome. Bender *et al.* (1983*a*) noticed that mutations of the Bithorax complex which are suppressed by *su(Hw)* are associated with *gypsy* insertions, and Modelell *et al.* (1983) suggested that this might be true of all mutations affected by *su(Hw)*. They tested this by hybridizing a *gypsy* probe to chromosomes carrying mutations suppressible by *su(Hw)*. They scored 14 mutations at eight loci, and in all but two cases were able to find hybridization to the chromosome position corresponding to the gene in question. The exceptions were two alleles of the *rudimentary* gene which are fully suppressed by *su(Hw)* at 18 °C but only partially suppressed at 25 °C. These could be associated with insertions of *gypsy* LTRs, which might not have been detected in this experiment. Alternatively, suppression may be independent of *gypsy*. These results support the suggestion that most, if not all, mutations suppressible by *su(Hw)* are associated with *gypsy* insertions. They do not indicate what proportion of mutations caused by *gypsy* insertions is suppressible by *su(Hw)*.

The effect of *su(Hw)* has been investigated for *gypsy* insertions in the *forked, f, and Hairy-wing, Hw*, genes. Accumulation of transcripts believed to be required for the wild-type *forked* phenotype is reduced in pupae carrying the f^1 mutation (Parkhurst and Corces 1985). This allele has a *gypsy* element inserted within the DNA coding for these transcripts, but it is not known whether this is within an intron or exon. These RNAs return to near normal levels in pupae carrying both f^1 and *su(Hw)*. This is also true of f^1 pupae carrying the *su(f)* mutation, another extragenic suppressor.

Parkhurst and Corces (1985) suggest that the *gypsy* element associated with f^1 affects transcription from the *forked* promoter, possibly because of an enhancer within the *gypsy* LTRs. The wild-type *su(Hw)* product may

interact with a regulatory element in *gypsy* to increase its effect on *forked*. The simplest interpretation would be that this product stimulates *gypsy* transcription and as a result *forked* transcription is reduced. The developmental pattern of *gypsy* expression is not affected by the *su(Hw)* or *su(f)* mutations, but this is not necessarily at odds with this model (Parkhurst and Corces 1985).

The *Hairy-wing* mutation, Hw^1, is expressed by *su(Hw)*, and is associated with a *gypsy* insertion in the *achaete* part of the *achaete–scute* complex (Campazano *et al.* 1986). It is a dominant mutation and results in an increased level of transcripts from this region. In this case the transcripts are truncated within the *gypsy* insertion. They return to near wild-type levels in the presence of *su(Hw)*, again suggesting that the $su(Hw)^+$ product can modulate transcription of *gypsy* and adjacent sequences. The opposite effects of *gypsy* elements on *forked* and *achaete* expression may be explained by their orientation. Transcription of the *gypsy* element in f^1 would be opposed to that of the *forked* gene, whereas transcription of the Hw^1 element would be the same as that of *achaete*.

The *gypsy*/*su(Hw)*, *su(f)* system is not the only interaction of this type in *D. melanogaster*. Searles and Voelker (1986) have found that alleles of *vermillion* which are strongly suppressed by the extragenic suppressor *su(s)* are associated with *412* insertions. They have shown by *in situ* hybridization, that alleles of *purple* and *speck* which are suppressed by *su(s)* may also be associated with *412* elements. The phenotypic affect of the *white–apricot* mutation is reduced by *suppressor of white–apricot*, $su(w^a)$, and increased by *su(f)* (Green 1959). This presumably reflects interactions between the products of the suppressor genes and the *copia* element in w^a (Levis *et al.* 1984; Pirotta and Brockl 1984). Similar effects have been found in other organisms. Some of the mutations due to *Ty* insertions in *Saccharomyces cerevisiae* are suppressed by mutations at other loci (Roeder and Fink 1983; Roeder *et al.* 1985), and the *dilute* mutation in mouse, which is associated with insertion of an endogenous retrovirus, is suppressed by the unlinked recessive mutation *dilute suppressor* (Copeland *et al.* 1983; Sweet 1983).

B. *COPIA*-LIKE ELEMENTS AND RETROVIRUSES

The structure of *copia*-like elements is very similar to that of the DNA proviruses of vertebrate retroviruses. Both have a central region several kb long flanked by LTRs of a few hundred bases. This structural similarity, and the fact that both types of element can insert into chromosomal DNA in a semi-random manner, encouraged several groups to investigate possible relationships between them.

The retroviral genome is a long single-stranded RNA with short direct repeats at each end. After a virus has infected a cell, this single-stranded RNA is converted to double-stranded DNA by the action of reverse transcriptase

incorporated in the virus particle. Synthesis of the first DNA strand is primed by a specific tRNA, also present in the virus particle. The last 18 bases of this tRNA pair with a unique sequence, the primer binding site, PBS, near the 5′ end of the viral RNA. It is not certain how synthesis of the second strand is primed, but it is believed to require a purine-rich sequence found near the 3′ end of the viral RNA. The molecules generated by reverse transcription are linear double-stranded DNAs with LTRs at each end. These repeats are composite structures containing sequences from three different parts of the viral genome (Fig. 1.2), a unique sequence from the 3′ end, U3, one copy of the sequence repeated at both ends of the viral RNA, R, and a unique sequence from its 5′ end, U5. This process is discussed in detail by Varmus (1983) and in Weiss *et al.* (1984). Some of this linear double-stranded DNA is transported to the nucleus where it circularizes to give two predominant forms. Both contain all the DNA between the LTRs of the linear molecule, but one has a single LTR while the other has two LTRs in tandem. Some of these circular DNAs then integrate into the chromosomes of the infected cell to generate the proviral genome which resembles linear precursor DNA in having LTRs at both ends. The life-cycle of the virus is completed by synthesis of genomic RNA, starting at the beginning of the R sequence in one LTR and terminating at the end of R in the other.

Two lines of evidence suggest that integration takes place at the junction of the tandem LTRs. Molecules containing tandem LTRs can integrate into chromosomes at this junction (Panganiban and Temin 1984*a*), and integration requires a viral encoded endonuclease which cuts preferentially at this site (Duyk *et al.* 1983). The LTRs at the ends of cytoplasmic linear DNA are believed to be four bases longer than those of integrated provirus (Scott *et al.* 1981). These four bases separate the tandem LTRs in the large circular molecules. Two of them are encoded immediately 5′ to the PBS while the other two are present just 3′ of the purine-rich sequence. These bases are lost during integration.

The similarity between *copia*-like elements and retroviruses is far greater than simply the presence of LTRs. All *copia*-like elements for which there is sequence information, except *mdg*3, have purine-rich sequences adjacent to their right-hand LTRs, and sequences similar to primer binding sites following their left-hand LTRs. The LTRs themselves contain likely promoter and polyadenylation signals, and in many cases start with the sequence TG and end CA (see appendix) as do all known retroviral LTRs (Weiss *et al.* 1985). The PBSs of *412*, *mdg*1, *297* and *17.6* are similar to that of avian leukosis virus (Kugimya *et al.* 1983), while that of *B104* is similar to that of Moloney sarcoma virus (Scherer *et al.* 1982).

The full extent of the relationship between *copia*-like elements and retroviruses has come to light with the publication of the complete nucleotide sequences of one *17.6* (Saigo *et al.* 1984) and two *copia* elements (Emori *et al.* 1985; Mount and Rubin 1985). The central regions of most retroviruses have

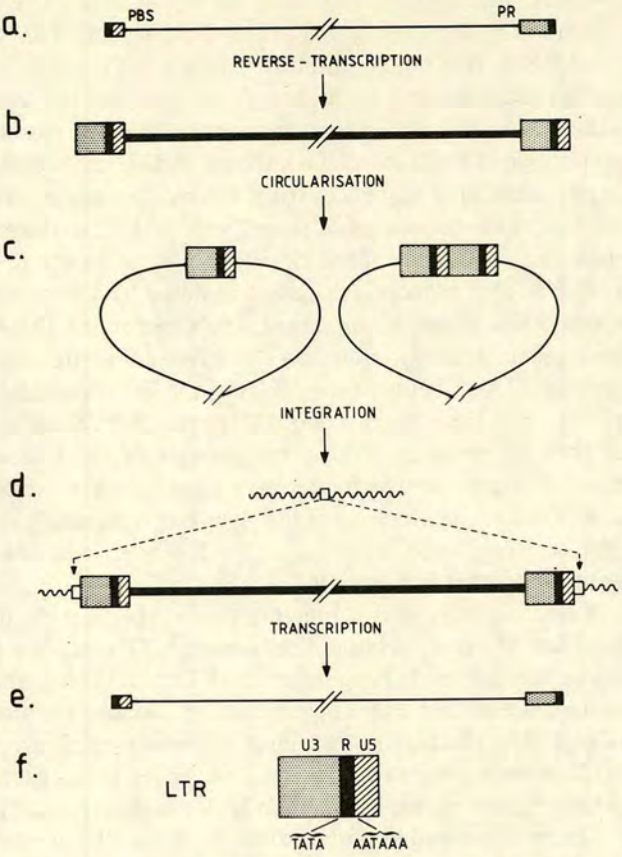


Fig. 1.2 Intracellular life-cycle of a retrovirus, and likely mechanism of transposition for *copia*-like elements. The structure of genomic viral RNA is shown in (a). The filled boxes indicate a sequence repeated at both ends of the RNA. The hatched and stippled boxes indicate unique sequences, U5 and U3, from the 5′ and 3′ ends of the RNA, respectively. This RNA is reverse transcribed into extrachromosomal linear double-stranded DNA, (b). Synthesis of the first DNA strand requires a tRNA primer bound to a PBS adjacent to U5. Synthesis of the second strand is thought to require a purine-rich sequence, PR, adjacent to U3. Linear DNA formed in this way has long terminal repeats, LTRs. These are made up of the sequences U3, R and U5 as shown in (f). These linear molecules can circularize to give molecules with a single LTR, or with two LTRs in tandem, (c). Circular molecules with two LTRs can integrate into chromosomal DNA, and in doing so duplicate a short target site sequence indicated by open boxes, (d). LTRs contain promoters, TATA, and polyadenylation signals, AATAAA, (f), allowing the integrated provirus to direct synthesis of full-length genomic RNA to start another cycle, (e). *Copia*-like elements could transpose as the result of a similar series of events, starting with full-length RNA.

three open reading frames known as *gag*, *pol* and *env*, each of which codes for polyproteins. The *gag* gene product is cleaved into four or five products which make up most of the core of the virus particle. One of the products of the Rous sarcoma virus (RSV) *gag* protein, p15, has protease activity and is probably involved in cleavage of viral gene products. It is encoded at the 3′ end of *gag*. The *pol* gene product, reverse transcriptase, is translated as a large molecule fused to the *gag* polypeptide. This results from a small splice, a translational frameshift or nonsense suppression. The polymerase is then processed from this large precursor. The reverse transcriptases of avian leukosis-sarcoma viruses (ALSV) are dimeric enzymes, one subunit of which can be cleaved to give a C-terminal polypeptide, p32, with endonuclease activity. The N-terminal end remains part of the polymerase. This nuclease is believed to cut at the junction between two tandem LTRs in circular viral DNA as described above (Duyk *et al.* 1983). It must play a role in integration since this is blocked by mutations near the 3′ end of the *pol* gene (Donehower and Varmus 1984; Panganiban and Temin 1984*b*; Schwartzberg *et al.* 1984; Panganiban 1985). The third open reading frame in the viral genome, *env*, overlaps the end of *pol* and is translated from a spliced sub-genomic message. It codes for another polyprotein which is processed into several membrane proteins forming major constituents of the viral envelope. More details on the structure of the retroviral genome can be found in Varmus (1983) and Weiss *et al.* (1984).

The sequence of the *17.6* element has several features in common with a retroviral genome. It contains three open reading frames (ORFs) which overlap slightly. The first, ORF1, has some amino acid sequence homology with the *gag* product of Moloney leukaemia virus suggesting that they may be equivalent. More striking similarities can be seen between ORF2 of *17.6* and *pol* of ALSV. Patarca and Heseltine (1984) have compared the reverse transcriptases of several retroviruses, cauliflower mosaic virus and hepatitis B virus, and have found regions of similar amino acid sequence. The ORF2 of *17.6* has considerable homology to a region believed to be important for polymerase activity, and there is also homology with the p32 region of ALSV and Moloney murine leukaemia virus (MoMLV) (Fig. 1.3). The relative positions of the putative polymerase and endonuclease regions in *17.6* are the same as those of their counterparts in MoMLV and ALSV (Fig. 1.4). No amino acid sequence similarities have been detected between ORF3 of *17.6* and the retroviral *env* genes.

Copia elements are about 2 kb shorter than proviruses and most other *copia*-like elements. The entire base sequences of the two different *copia* elements have been determined (Emori *et al.* 1985; Mount and Rubin 1985). They are remarkably similar and differ by only four silent single-base substitutions in coding regions, and three single-base deletion/insertions in the 3′ untranslated region. These sequences contain a single ORF of 4227 nucleotides. The N-terminal region of this ORF has some similarities to *gag* genes,

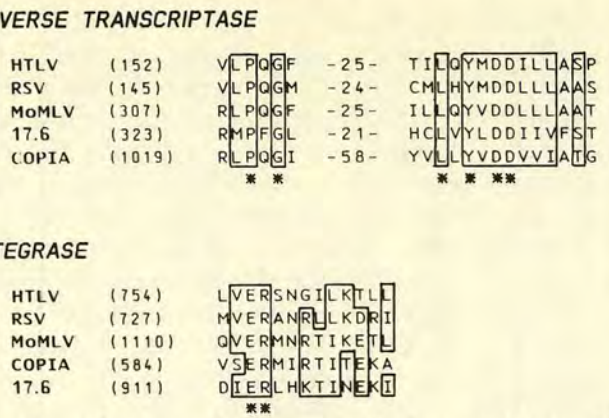


Fig. 1.3 Comparison of amino acid sequences coded by retroviruses and *copia*-like transposable elements. The sequences shown are believed to be associated with reverse tanscriptase or integrase activity. The sequences are as follows: HTLV, human adult T-cell leukaemia virus (Seiki *et al.* 1983); RSV, Rous sarcoma virus (Schwartz *et al.* 1983); MoMLV, Moloney murine leukaemia virus (Schinnick *et al.* 1981); *copia* (Mount and Rubin, 1985; Emori *et al.* 1985); *17.6* (Saigo *et al.* 1984). The numbers in brackets indicate the distance from the start of the appropriate protein or ORF. The numbers not in brackets indicate the number of residues separating the two sequences from the reverse transcriptase-like polypeptides. Positions at which all five elements have identical, or similar, residues, are boxed. Asterisks indicate positions with invariant residues. Amino acid residues are indicated as follows; A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; Y, tyrosine. Amino acids were grouped as follows for making these comparisons: P,A,G,S,T—neutral or weakly hydrophobic; Q,N,E,D—hydrophylic, acid amine; H,K,R—hydrophylic, basic; L,I,V,M—hydrophobic; F,Y,W—hydrophobic, aromatic; C—cross-link forming.

and in particular to a region of *gag* believed to code for a nucleic acid binding protein. The amino acid sequence downstream of this putative nucleic acid binding domain is similar to that of the p15 region of RSV, and contains the tripeptide Asp-Ser-Gly found in the active site of trypsin proteases (James 1980).

Shiba and Saigo (1983) have found particles that morphologically resemble retroviral core particles, in *D. melanogaster* tissue culture cells. These particles contain a heterogeneous population of RNAs including a prominent 5 kb species complementary to *copia* DNA. Emori *et al.* (1985) have determined a composite cDNA sequence for this RNA. It is equivalent to that of *copia* DNA and differs by only a few bases. One base is missing near the end of the long ORF of *copia* DNA and as a result the cDNA sequence contains two ORFs. This may indicate that most of the *copia* RNA from which the cDNA was made had been transcribed from a defective *copia* element(s). The cDNA sequence starts within the left-hand LTR and, like

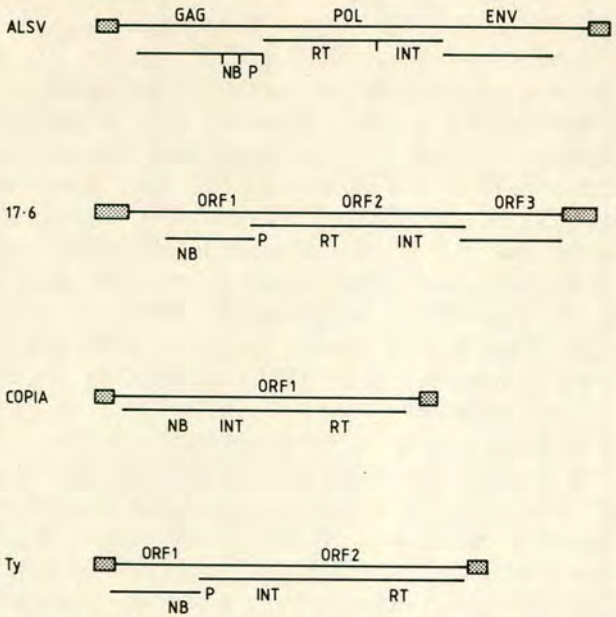


Fig. 1.4 Comparison of the genomes of an avian leukosis sarcoma virus, ALSV, the *copia*-like elements *17.6* and *copia*, and the yeast transposable element Ty. The boxes indicate LTRs. ORFs are indicated by horizontal lines below each map. The *gag*, *pol* and *env* genes of ALSV are described in the text. The region of the ALSV genome coding for amino acid sequences with particular properties are indicated as follows: NB, nucleic acid binding; P, protease; RT, reverse transcriptase; INT, integrase. The regions of the transposable elements believed to be associated with similar functions are indicated in the same way.

cellular *copia* RNA, has variable 5′ ends. The 3′ end of the sequence is within the right-hand LTR.

The RNA from these virus-like particles can be translated inefficiently in a heterologous *in vitro* system to give several polypeptides including a prominent 31 000 u species (Shiba and Saigo 1983). This appears to be a structural component of the virus-like particles since it reacts with antisera raised against the particles themselves. Emori *et al.* (1985) have reported that a partial amino acid sequence of the 31 000 u translation products corresponds to that of the N-terminal region of the *copia* ORF. This would include the putative nucleic acid binding domain noted by Mount and Rubin (1985), consistent with the idea that this is a *gag*-like region.

The C-terminal segment of the *copia* ORF has regions which, like ORF2 of *17.6*, have some homology with the reverse transcriptase and endonuclease domains of retroviral *pol* genes (Fig. 1.3), but in this case their relative positions are reversed (Mount and Rubin, 1985) (Fig. 1.4). The product of this long ORF may be cleaved into *gag* and *pol*-like regions by the protease activ-

ity of the p15-like sequence. The *copia* sequence does not have any region which could correspond to *env* or ORF3 or *17.6*.

This sequence organization is not unique to *copia* elements. The transposable *Ty* elements in *S. cerevisiae* are very similar to *copia* elements (see review by Roeder and Fink 1983). They are about the same size and have LTRs at both ends. Clare and Farabaugh (1985) have determined the base sequence of one *Ty* element *Ty912*. It has two ORFs which overlap slightly. Segments of the first ORF are reminiscent of DNA binding proteins, while the second shows homology to the protease, reverse transcriptase and endonuclease regions of retroviral *pol* genes, and is partially homologous with the corresponding regions of *copia* (Mount and Rubin 1985) (Fig. 1.3). Again there is no region equivalent to *env* or ORF3 of *17.6*. *Copia* elements are thus more like *Ty* elements in yeast than *17.6* elements in *D. melanogaster* even though all three are clearly similar to retroviruses (Fig. 1.4).

Long before the coding capacities of *copia*-like elements and retroviruses were known in detail several authors suggested that these elements might transpose by similar mechanisms. The DNA–RNA–DNA cycle of a retrovirus provides a means of transposition without recourse to extrachromosomal viral particles. This is almost certainly true of *Ty* elements. Boeke *et al.* (1985) have followed the behaviour of a marked *Ty* element containing an intron. This element was able to transpose, and in doing so regenerated an element lacking the intron, as would be expected if transposition were to take place *via* a truncated retroviral life-cycle.

There is, as yet, no direct evidence that the same is true of *copia*-like elements, although this seems very likely. Circular molecules containing *copia*, *412*, *mdg*1, *mdg*3, *mdg*4/*gypsy* and *297* elements have been found in extrachromosomal DNA from *D. melanogaster* tissue culture cells and embryos (Flavell and Ish Horowicz 1981; Ilyin *et al.* 1984; Junakovic and Ballario 1984; Shepherd and Finnegan 1984; Mossie *et al.* 1985). The number of circular molecules per cell, for any particular element, depends on the cells being tested. The average number is about one or less per cell, but there may be considerable variation between cells. Circular *B104*/*roo* and *HMS Beagle* molecules have been looked for in tissue culture cells but were not found (Junakovic and Ballario 1984; Mossie *et al.* 1985).

Flavell and Ish Horowicz (1983) have cloned individual circular *copia* molecules isolated from K_c tissue culture cells (Echalier and Ohanessian 1969). The majority contained all the sequences from the body of a *copia* element, plus one LTR or two LTRs in tandem. The latter molecules seem to be peculiar to K_c cells since they have not been found in two other cell lines (Ilyin *et al.* 1984, Mossie *et al.* 1985). Shepherd and Finnegan (1984) have cloned circular *412* molecules from the K_c line and have found that, in this case, the majority of molecules contained only a single LTR. They could detect no molecules with two LTRs in tandem. Both groups found circular molecules with deletions, inversions or other rearrangements.

Circular *copia*-like elements could be formed in several ways. Recombination between the LTRs at the ends of a chromosomal element will generate circular molecules with a single LTR, while recombination between the short direct repeats flanking an element will generate molecules with two LTRs in tandem. The LTRs in *copia* circles of this type should be separated by 5 bp since this is the length of the target site duplication characteristic of these elements. Flavell and Ish Horowicz (1983) have sequenced the junctions between the tandem LTRs of several cloned circular molecules. The number of bases separating them varied between 0 and 15 bp, but in no case was it 5 bp. If these molecules had been formed by excision of chromosomal elements, then these must have been aberrant excision events.

Circular molecules could also be formed from linear DNA produced by reverse transcription of full-length RNA. Flavell (1984) has used density labelling techniques to show that many circular *copia* molecules incorporate label faster than the bulk of chromosomal DNA and are probably not produced by recombination. He grew K_c cells in medium containing bromodeoxyuridine, and compared the rate at which this was incorporated into chromosomal and extrachromosomal elements. Circular elements incorporated the label into both strands more rapidly than did chromosomal elements. He was also able to find full-length linear *copia* molecules in extrachromosomal DNA which had incorporated label in both strands. The formation of these linear molecules was not affected by an inhibitor of the major cellular DNA polymerase, consistent with the idea that they were formed by reverse transcriptase.

Ilyin *et al.* (1984) have found circular *mdg*1 and *mdg*3 elements in two cell lines, and in one of these Arkhipova *et al.* (1984) have identified *mdg*1 and *mdg*3 molecules in the form of RNA/DNA hybrids. The structure of these molecules suggested that they had been formed by reverse transcription.

Whatever the origin of circular *copia*-like elements it is unlikely that many of them are transposition intermediates. Circular molecules with two LTRs in tandem have been found only for *copia* elements in one cell line, and even then their structure was not that expected for retroviral circles. These data do not rule out the possibility that *copia*-like elements transpose by a retroviral mechanism. The frequency of transposition is usually very low (Rubin *et al.* 1981; Tchurikov *et al.* 1981) and transposition intermediates may be correspondingly rare. This is not necessarily at odds with the high levels of full-length *copia* RNAs. Many of these transcripts may come from defective elements, and any or all of the subsequent steps in transposition may be inefficient. These would include primer binding, reverse transcriptase synthesis and activity, formation of linear and circular DNAs, and integration of the appropriate precursor. Mossie *et al.* (1985) have found that the abundance of circular *copia*-like elements is not related to the levels of the corresponding RNAs. The amplification of *copia*-like elements in tissue culture cells is probably not related to high levels of transcription in these cells but may reflect

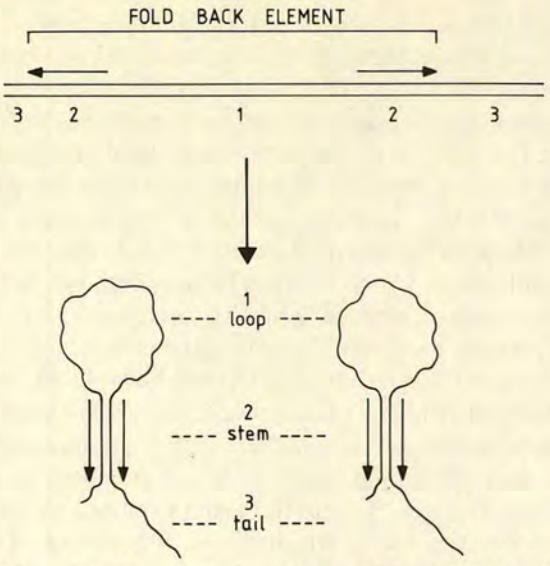


Fig. 1.5 Formation of fold-back structures as a result of denaturation and renaturation of a fragment containing an inverted repeat sequence. The regions of genomic DNA which form the loop, stem and tails of the fold-back structure are designated 1, 2, and 3, respectively.

reduced selection against high copy number as compared with the whole organism.

Boeke *et al.* (1985) have been able to increase the level of *Ty* transposition one-hundred fold by increasing the level of expression of a particular *Ty* element. This was done by fusing it to an inducible promoter. The frequency of transposition to a particular site was still low, of the order of one in 10^6 cells, but the overall level of transposition was high and could be detected easily without selection. Similar experiments with *copia*-like elements should allow the structure of newly transposed elements to be compared with the donor element, and might raise transposition intermediates to a detectable level. The difficulty with such an experiment is to be certain of finding a functional element to manipulate.

III. Transposable elements with long inverted repeats

A. FOLD-BACK ELEMENTS

All eukaryotic genomes so far tested contain closely spaced inverted repeat sequences. These can be detected by denaturing high molecular weight DNA and allowing it to reanneal at low concentration. The first sequences to renature will be inverted repeats. The structures formed in this way are usually referred to as 'snap-back' or 'fold-back' molecules (Fig. 1.5). These contain a double-stranded stem comprising the inverted repeats themselves, a

single-stranded loop containing the sequence separating the inverted repeats and single-stranded tails containing adjacent DNA. There are estimated to be 2000–4000 inverted repeat structures in the *D. melanogaster* genome, and DNA from all frequency classes are represented in their stems, loops and tails. About 3 per cent of the mass of the genome is contained in the stems which are slightly enriched for moderately repetitive sequences (Schmid *et al.* 1975).

Potter and his colleagues have studied one family of fold-back elements, the FB family, in detail. Potter *et al.* (1980) prepared a fraction of the genome enriched for inverted repeat sequences and used this as a probe with which to screen a library of genomic DNA cloned in a plasmid vector. One of the plasmids hybridizing to this probe, pDmFB1, contained a complete inverted repeat structure, FB1, and was used to isolate further clones containing related sequences. Nine fold-back elements, FB1–FB9, have been isolated in this way. They are related in sequence but, unlike members of a family of *copia*-like elements, are very heterogeneous in structure. The length of the inverted repeats and the length and sequence of DNA separating them differ from element to element, and some have no loop region at all (Potter *et al.* 1980; Truett *et al.* 1981). The sequence of these inverted repeats is present at the chromocentre, and about 20–30 sites on the arms of salivary gland chromosomes, although they do not necessarily occur as inverted repeats at each site. The exact locations of these sequences varies from one strain to another, suggesting that FB elements are transposable (Levis *et al.* 1982).

The inverted repeats themselves are made up of tandemly repeated short sequences in a pattern reminiscent of satellite DNA. This was first indicated by the pattern of restriction sites within the inverted repeats. Most enzymes do not cut at all, whereas TaqI cuts frequently with many of the sites being regularly spaced 155 bp apart (see appendix). The outer ends of the inverted repeats are highly conserved and contain the only HinfI sites within the repeats (Truett *et al.* 1981). Potter (1982*a*) has determined the entire base sequence of FB4, a fold-back element with a 1.7 kb loop region, and Truett *et al.* (1981) have obtained information for parts of FB3. The bulk of the inverted repeats are made up of short repeat sequences with some variation from one copy to another. Near the outer ends of an element this repeat is 10 bp long. This is expanded first to a 20 bp repeat present in multiple copies separated by variable A + T rich regions, and then to a 31 bp sequence which is repeated many times in tandem. There are five main variants on this 31 bp theme, and these occur in a cyclical pattern giving a larger repeat unit of 155 bp containing a single TaqI site. The same pattern has been found in the partial sequence of FB3 (Truett *et al.* 1981).

There are more copies of the 31 bp repeat in the right-hand portion of FB4 than in the left. Restriction mapping data from other elements suggest that this is often the case, and in many instances the loop region may be made up entirely of excess repeat sequences.

All but 95 bp of the loop of FB4 is bounded by nearly perfect inverted repeats 33 bp long. The sequence between these repeats is repeated about 20 times in the genome, and Brierley and Potter (1985) have called these HB elements, the HB element in FB4 being HB1. These elements appear to be a fairly stable component of the genome since very similar patterns of hybridization are seen when digests of genomic DNA from different strains are probed with HB DNA in a Southern transfer experiment. Brierley and Potter (1985) have isolated 23 plasmids containing HB DNA. None of these hybridizes to the inverted repeat sequences of FB4, indicating that HB elements are rarely part of FB elements. They have compared four HB elements, including HB1, and have found that they are poorly conserved but are all flanked by related inverted repeats 29–33 bp long. Brierley and Potter (1985) think that HB sequences are transposable elements in their own right because they have terminal inverted repeats, they contain moderately repeated sequences, and they are not usually part of FB elements. They suggest that FB4 may have been formed by transposition of HB1 into an FB element.

These data are indicative, but not conclusive. The similar patterns of hybridization of HB probes to genomic DNA from different strains indicates a low frequency of transposition. The differences which are seen could be due to sequence variation within, or adjacent to, HB elements, or to transposition of FB4 itself. More convincing evidence might be obtained from *in situ* hybridization experiments or by direct demonstration of transposition of an HB element into, or close to, a gene. No HB element has been found with flanking direct repeats as might be expected of a transposable element.

An FB element has been found associated with *white–crimson*, w^c, an unstable derivative of the mutation *white–ivory*, w^i (Green 1967; Collins and Rubin 1982). The w^i mutation is due to tandem duplication of 2.9 kb of *white* DNA, while w^c has a 10 kb FB element, FBw^c, inserted near the junction of the two repeats (Collins and Rubin 1982; Levis *et al.* 1982; O'Hare *et al.* 1984). Somewhat surprisingly the w^c allele determines a darker eye colour than does w^i. Collins and Rubin (1983) have analysed FBw^c in detail. It is flanked by a 9 bp target site duplication, like FB3, and the 31 bp at the outer ends of its inverted repeats are identical to each other, and to the termini of FB3 and FB4.

The w^c mutation produces w^+ or w^i derivatives with a frequency of about 10^{-3} (Green 1967), and Collins and Rubin (1982, 1983) have examined the structure of some of these. The w^+ derivatives have lost one copy of the duplicated *white* DNA plus FBw^c, presumably as the result of recombination between the 2.9 kb repeats. The w^i derivatives retain the *white* duplication, but have lost FBw^c. This has excised precisely since all FB sequences have gone, together with one copy of the 9 bp target site duplication. Precise excision of FBw^c is probably stimulated by the FB element itself since it occurs more frequently than would be expected for recombination between 9 bp direct repeats. The inverted repeats of the FB element may occasionally

pair with each other to bring the 9 bp direct repeats into close proximity. Alternatively an FB-related function might stimulate recombination between flanking sequences. Recombination between the 2.9 kb repeats of *white* DNA also occurs more frequently in chromosomes carrying the w^c mutation than in those carrying w^i. Again this could be stimulated by an FB-related function, or the presence of FBw^c may simply allow these repeats to pair more easily.

Not all DNA rearrangements stimulated by FB elements are simple insertions or excisions. The w^c mutation occasionally produces derivatives which determine a bleached white phenotype. Collins and Rubin (1984) have analysed 10 of these. Six were deletions of DNA to the left of the insertion, and could be explained by recombination between FBw^c and a second FB element lying about 14 kb to the left of it on the w^c chromosome. Three of the four remaining *white* alleles had small rearrangements of sequences within FBw^c. It is not clear how these were formed. The altered FB elements may have been produced from FBw^c as a result of the activity of a protein normally involved in transposition. Alternatively, they may be the result of recombination/gene conversion within FBw^c, or between FBw^c and other FB elements elsewhere in the genome. This would not necessarily change the loop sequence since this is repeated a few times in the genome and is usually associated with FB sequences (Levis *et al.* 1982; Paro *et al.* 1983; Brierley and Potter 1985).

These rearrangements are not peculiar to FBw^c. Potter (1982*b*) has found a similar change in the FB4 element. This could also be the result of either intra-element events, or recombination/gene conversion between elements.

The frequency with which FB elements can excise seems to be affected by their structure. Collins and Rubin (1984) have found a stable derivative of FBw^c which apparently differs from it only in having a duplication of the central region. Its stability was not due to lack of any destabilizing product coded by FBw^c, since the derivative was not affected by the presence of a second, unstable, FBw^c.

Nothing is known about the mechanism of transposition of FB elements nor what proportion of these elements are transposition proficient. The structure of FB elements suggests that they do not transpose *via* an RNA intermediate. The fact that FBw^c can excise precisely does not necessarily mean that excision is part of transposition.

B. TE ELEMENTS—COMPOSITE TRANSPOSABLE ELEMENTS

Fold-back elements can form part of compound transposable elements. The first of these to be discovered, the TE elements, were detected genetically by Ising and Ramel (1976). They found a strain of *D. melanogaster* in which the *white* and *roughest* genes, which normally reside close together on the X chromosome, had moved to the second chromosome. They were not stable at this

new site but transposed again, or were apparently lost altogether, at a frequency of about 10^{-3} (Ising and Block 1981). The amount of DNA transposing must be several hundred kilobases, and in favourable cases several polytene chromosome bands can be seen inserted at the site of a TE (Ising and Block 1981, 1984). The element responsible for the first transposition event is called TE1. About 150 transpositions have been identified, the TE element being renumbered after each event, and about 90 have been mapped cytologically (Ising and Block 1984). TE elements can change their genetic content, and this is often associated with transposition. They have been found to have a near wild-type allele, instead of the *white–apricot* allele carried by TE1, to have duplicated or lost the entire *white* gene, or to have incorporated new genes (Ising and Block 1981, 1984; Paro *et al.* 1983).

Paro *et al.* (1983) have examined both ends of TE77 and TE98 and one end of TE28. Both TE77 and TE98 have FB elements at each end. The FB elements at the two ends of both TE77 and TE98 are not identical. There is also an FB element at the end of TE28, which was examined. This has been called FB-NOF, and appears to be the same as FBw^c (Paro *et al.* 1983).

Although the genetic markers carried by TEs may be lost, the TEs themselves probably do not excise precisely. Ising and Ramel (1976) have found several examples of transpositions of a TE element being associated with induction of a recessive mutation at, or near to, the site of insertion. These mutations may remain even after the *white* gene has been lost. This suggests that part of a TE element can be lost without the element excising precisely. An alternative, but less likely, explanation is that these mutations arose independently of TE insertion.

Chia *et al.* (1985) have studied excision of TE146 at the molecular level. This element carries two copies of the *white* gene and is inserted near the *no-ocelli* gene, *noc*, resulting in a noc^- mutation. Chia *et al.* (1985) selected chromosomes which had lost both the w^+ alleles of TE146. They cloned DNA spanning the insertion site of this element and used it to probe DNA from these derivatives in Southern transfer experiments. All eight chromosomes tested had DNA inserted at the TE146 target site, and *in situ* hybridization experiments indicated that at least some of this was FB sequences. They interpret these results as indicating that most of TE146 has excised as a result of recombination between inverted repeat sequences in the FB elements at the ends of TE146. The loss of the w^+ alleles was accompanied by reversion of the noc^- mutation, suggesting that TE146 had inserted adjacent to, rather than within, the *noc* gene.

The loss of the bulk of TE146 occurred at a frequency of about 5×10^{-5} (Gubb *et al.* 1986) whereas loss of only one copy of w^+ was about 10 times more frequent. Chia *et al.* (1985) suggest that this is the result of recombination between one of the terminal FB elements and an FB element within TE146. These derivatives are still noc^- but throw off noc^+ chromosomes which are indistinguishable from the single-step revertants described above.

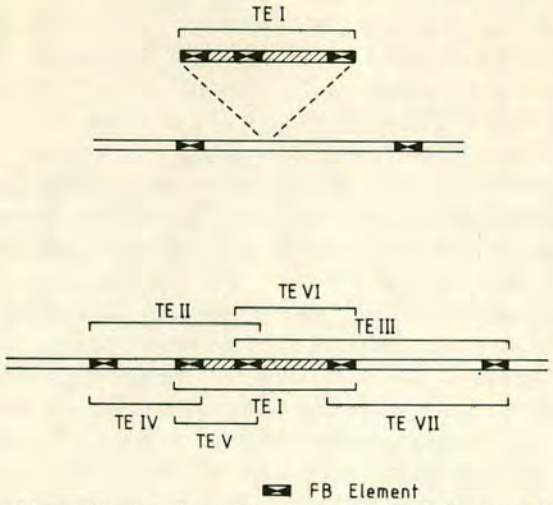


Fig. 1.6 A possible mechanism for generation of new TE elements. An element, TEI, containing an internal FB element, has inserted between two chromosomal FB elements. Transposition events could generate six new TE elements. TEII–TEVII, each containing at least part of TEI.

The presence of FB elements within a TE element could explain how their genetic constitution changes during transposition. If a TE element, TEI, containing an internal FB element, transposes to a site flanked by other FB elements, as will often be the case, then several new TEs could be formed by subsequent transpositions using different combinations of FB elements (Fig. 1.6). Two of the new TEs would also have internal FB elements. One, TEII, would have the FB element within TEI at its right-hand end, and the other, TEIII, would have this at its left-hand end. Four TEs, TEIV–TEVII, could be formed without internal FB elements, and there are more complex possibilities. The probability of any particular structure being formed may depend on the distance separating FB elements, although some FB elements may be unable to transpose themselves and/or to form an end of a TE.

Another transposable element flanked by FB sequences has been found inserted just upstream of the start of the *white* gene (Bingham 1980, 1981; Levis *et al.* 1982; O'Hare *et al.* 1984), and is responsible for the unstable *white* mutation w^{DZL}. This element is 13 kb long and has FB elements surrounding a 6.5 kb sequence. The DNA of this central region is not repeated in the genome and comes from near the tip of the second chromosome (Levis *et al.* 1982). This element is unstable and can rearrange to give w^{+} derivatives with a frequency of 10^{-2}–10^{-3} (Bingham 1981), the majority of which are simple in that they are not associated with gross chromosomal rearrangements.

Levis and Rubin (1982) have analysed the *white* locus in 12 simple revertants, 11 of which retain some DNA inserted at the site of the w^{DZL} element. Most of the residual insertions are less than 4 kb long and appear to have lost

all of the central region of the w^{DZL} element. The remaining DNA presumably includes FB sequences, and could have been formed by recombination between the inverted repeats at the ends of the w^{DZL} element. Many of these w^+ derivatives are themselves unstable and throw off mutant *w* alleles. This is consistent with there being FB sequences near the *white* gene.

TE elements almost certainly transpose because of the FB elements at their ends, and possibly any sequence flanked by FB elements is potentially transposable. The size and structure of TE elements make it unlikely, if not impossible, that they transpose *via* an RNA intermediate. No one has compared the structure of a TE element before and after transposition to determine whether or not its terminal FB elements are altered, nor has anyone examined a target prior to TE insertion. All the rearrangements associated with TE instability, and some of those due to FB elements, can be explained by recombination events involving sub-repeats of the FB inverted repeats. These could take place within or between FBs, depending on the nature of the rearrangement concerned. Potter (1982) has suggested that these sub-repeats bind a transposase, and recombination between sub-repeats might be stimulated by the same enzyme. It is not inconceivable that transposition of TEs is the result of recombination between the FB elements at the ends of a TE, and an FB element already at the target site. We have already pointed out that the sequence organization within the inverted repeats of an FB element is like that of satellite DNA. Perhaps FB elements evolved from these highly repetitive sequences, with variants of a satellite DNA binding protein becoming a transposase.

IV. F elements—transposable elements without terminal inverted repeats

The F elements of *D. melanogaster* are disperse repeat sequences which resemble retroposons (Rogers 1983), a class of transposable sequences found in many other species. These include Alu sequences in human DNA (Houck *et al.* 1979; Jelinek *et al.* 1980) and the corresponding short repeated sequences in rodents (Kramerov *et al.* 1979; Krayev *et al.* 1980; Haynes *et al.* 1981), the long interspersed sequences, or LINES, in mammalian genomes (Singer and Skowronski 1985), pseudogenes complementary to rat and human small nuclear RNAs (van Arsdell *et al.* 1981) and human tubulin and immunoglobulin pseudogenes (Hollis *et al.* 1982; Wilde *et al.* 1982). These sequences are flanked by target site duplications but are distinguishable from the transposable elements we have discussed so far in that they have no terminal repeats, either direct or inverted. The length of the target site duplications varies from one copy of an element to another, and there is an (A)-rich tract at the 3′ end of one strand. This is conventionally taken as the right-hand end of an element. Sequences of this type have been discussed in detail by Rogers (1985).

The first F element to be discovered was found inserted into one of the non-nucleolar copies of the type I sequence found inserted in many 28 S rRNA genes (Dawid *et al.* 1981). Dawid *et al.* (1981) suggested that this was a transposable element as it was flanked by direct repeats of the target sequence. This element, called 101F, is 4.7 kb long. It has no terminal repeats but has an 18 base polyA sequence at the end of one strand. This is preceded by two overlapping copies of the polyadenylation signal AATAAA. There are about 50 copies of sequences related to 101F in the *D. melanogaster* genome. About half of these are on the chromosome arms, and half in centromeric DNA. Dawid *et al.* (1981) and Pardue and Dawid (1981) have found very different sites of F element hybridization on chromosomes from different strains, which is consistent with the idea that these are transposable sequences.

Di Nocera *et al.* (1983) have compared the base sequence of 101F with that of three other F elements which they had cloned using 101F as a probe. All four elements are flanked by target site duplications which range from 8 to 13 bp long. All have polyA at the 3′ end of the same strand and in each case this is preceded by at least one polyadenylation signal. The left-hand ends of these elements are variable. One has a long deletion at this end while the others have small length differences and regions of non-homology. A fifth F element has been found inserted at the *white* locus in a chromosome carrying a derivative of the w^i mutation, w^{i+A} (O'Hare *et al.* 1984). This is similar to the others but has a large deletion at its left-hand end (Di Nocera *et al.* 1983).

The properties of F elements, and other retroposons, suggest that they may transpose by integration of double-stranded molecules made by reverse transcription of polyadenylated RNAs, although there is no direct evidence to support this. The short polyA sequence at their right-hand ends are taken as being relics of the terminal polyA, and the variability at their left-hand ends could be due to premature termination of reverse transcription. The polyadenylated RNAs used for transposition probably come from a few master copies of the element which are transcribed either because they are adjacent to RNA polymerase II promoters, or because they contain an internal polymerase III promoter. A transposed copy of the sequence will only be able to generate further transposition events if it has inserted adjacent to a suitable promoter, or if it carries with it an internal promoter. Neither is true of 101F, the longest F element so far described. This distinguishes F elements from *copia*-like and *Ty* elements, and retroviruses which probably also transpose *via* reverse transcription. These elements have polymerase II promoters within their LTRs, and can direct synthesis of RNAs from which full-length elements can be regenerated. The F elements found at variable sites in the genome might more correctly be described as transposed elements than transposable elements. Dawid *et al.* (1981) were unable to detect any $polyA^+$ transcripts complementary to F elements in embryos, larvae, pupae, and adults.

Di Nocera *et al.* (1983, 1986) have found another family of potential retroposons which they call G elements, which resemble F elements in that they have a polyadenylation signal and polyA tract at the 3′ end of one strand. They were found inserted in some of the F elements interrupting non-nucleolar copies of the 28S rDNA insertion sequence. There are 10–20 G elements in the genome. This interdigitation of repeated and possibly transposable sequences is not unique. Wensink *et al.* (1979) have described clusters of scrambled repeat sequences and Tchurikov *et al.* (1981) have found stretches of DNA containing several different dispersed repeat sequences.

V. Transposable elements in hybrid dysgenesis

A. HYBRID DYSGENESIS

Hybrid dysgenesis is the production of abnormal characteristics in the progeny when particular strains of flies are crossed in an appropriate fashion. These traits include partial or complete sterility, increased mutation frequencies, chromosome rearrangements, distorted transmission ratios and male recombination. There are two independent systems of hybrid dysgenesis, P-M and I-R. P-M dysgenesis is produced by crossing M, maternal, strain females with P, paternal, strain males. In the I-R system R, reactive, strain females must be crossed with I, inducer, strain males. The progenies of the reciprocal crosses are apparently normal in each case.

The biological properties of the P-M and I-R systems are not identical. The most obvious differences are in the nature of the induced sterility, and the fact that P-M dysgenesis affects both sexes, whereas I-R dysgenesis affects only female progeny. The gonads of both male and female progeny of a P-M dysgenic cross fail to develop. The gonads of I-R dysgenic females appear normal, and the flies lay normal numbers of eggs, but these embryos die at a very early stage of development. Dysgenic events are believed to be confined to the germ-line in both systems, since somatic mutations are rarely seen.

The characteristics of P strains and I strains are controlled by transposable elements called P factors and I factors. These elements are dormant until they are introduced into the cytoplasmic background of an M strain, in the case of P factors, or an R strain in the case of I factors. This is believed to be due to regulatory molecules which are produced by P or I factors, and which are not present in the nuclei of M or R strains. A detailed discussion of the genetic and biological aspects of hybrid dysgenesis can be found in Kidwell (1983), Bregliano and Kidwell (1983) and O'Hare (1985).

B. P-M DYSGENESIS AND P ELEMENTS

Many P-M induced mutations are unstable in P-M dysgenic individuals (Engels 1979; Rubin *et al.* 1982) and several people suggested that they might

be due to insertion of P factors into the genes in question (Golubovsky *et al.* 1977; Green 1977; Simmons and Lim 1980). This has been confirmed by Rubin and his colleagues (Rubin *et al.* 1982). They compared the structure of the wild-type *white* gene with that of seven P-M induced *white* mutations. Each mutation was associated with a DNA insertion in the *white* gene. Two of these insertions were of *copia* elements, while the other five were of sequences related to each other and ranging in size from 0.5–1.4 kb. When chromosomes carrying these mutations were subject to P-M dysgenesis, those associated with *copia* elements were stable whereas the others reverted to wild-type. Southern transfer analysis of these revertants suggested that the inserted sequences had excised precisely, and this has been confirmed subsequently by DNA sequencing (O'Hare *et al.* 1983). The elements associated with the unstable mutations were considered too short to be functional P factors, and Rubin *et al.* (1982) suggested that they might have been derived from complete P factors by deletion. O'Hare and Rubin (1983) searched for longer, related sequences using one of the short elements as a probe. They screened a library of clones P strain DNA and recovered several fragments containing a conserved 2.9 kb sequence.

Spradling and Rubin (1982) have confirmed that these longer elements are active P factors by showing that one of them could stimulate hybrid dysgenesis when injected into embryos of an M strain. They used an M strain carrying a mutation of the *singed* gene, sn^w, which had been induced by P-M dysgenesis and is extremely unstable in dysgenic flies (Engels 1979). If the injected DNA were to induce hybrid dysgenesis this would be revealed by mutation of sn^w to an almost wild-type allele, or a more extreme *singed* mutation. This was scored in the progeny of flies resulting from the injected embryos. Injection of the putative P factor did destabilize sn^w in this way. The P factor was itself affected by dysgenesis and transposed from plasmid DNA to chromosomes of the injected embryos. It integrated by transposition rather than recombination since the complete P factor inserted into the chromosomes without any flanking sequences. This has since proved to be an effective transformation system for *Drosophila* (Rubin and Spradling 1982).

The fact that two of the P-M induced *white* mutations were due to *copia* insertions does not necessarily mean that P-M dysgenesis can stimulate *copia* transposition. These insertions could have occurred fortuitously, or could have been stimulated by crossing the strains used in the experiment but not by P-M dysgenesis itself. P-M dysgenesis can only be invoked if *copia* transposition were to increase in the progeny of a cross between M strain females and P strain males, but not in the progeny of the reciprocal cross.

O'Hare and Rubin (1983) have determined the base sequence of one complete P factor and four of the shorter P elements isolated from the P-M induced *white* mutations. The P factor has 31 bp inverted repeats at each end, and is flanked by a target site duplication 8 bp long. The shorter P elements share these features, and differ from the 2.9 kb elements by internal deletions.

P elements do not transpose at random, and three of the five P elements in the *white* gene had inserted at exactly the same position. O'Hare and Rubin (1983) have compared the 8 bp site duplications of several P elements and have been able to deduce a weak consensus sequence from them. This may account for the target site specificity.

The genomes of most P strains contain 30–50 P elements distributed on all chromosomes (Bingham *et al.* 1982), only about 10 of which are complete (O'Hare 1985). The simplest explanation for the genetic differences between P and M strains would be that P elements are present in P strains but not M strains. This is true of some, but not all, M strains (Bingham *et al.* 1982). Many M strains, particularly those recently isolated from the wild, contain about as many P sequences as do P strains (Fig. 1.7a) (Kidwell 1983; Anxolabehere *et al.* 1984, 1985), but these are of genetically inactive P elements (Engels 1984). Strains of this type are sometimes called M′ strains. The fact that some strains of *D. melanogaster* lack P sequences altogether suggests that P elements have been incorporated into the genome relatively recently. Studies of the distribution of P sequences in other *Drosophila* species support this conclusion (Brookfield *et al.* 1984; Daniels *et al.* 1984).

One strand of the P factor contains four ORFs of 297, 714, 792 and 654 bp. Karess and Rubin (1984) have made mutations in each of these separately, and have tested the ability of the resulting mutant P factors to destabilize the sn^w mutation after being introduced into M strain embryos. Mutations in any one ORF blocked the trans-acting function of P required for transposition, but did not affect the ability of the mutant P element to transpose when coinjected with a complete P factor. These ORFs must determine a single function since the mutant P factors were unable to complement each other in any pair-wise combination. Presumably the information in these ORFs is joined by RNA splicing, and codes for a function required for transposition. Laski *et al.* (1986) have recently obtained evidence that this splicing may be confined to the germ cells of dysgenic individuals, explaining the tissue specificity of P-M dysgenesis.

Karess and Rubin (1984) have looked for P factor transcripts in polyA^+ RNA from P strain embryos, and from embryos obtained from a P-M dysgenic cross. They found RNAs ranging in size from 0.5 kb to more than 4 kb in each case, with a prominent band of 2.5 kb. Many of the shorter transcripts could come from incomplete P elements, while those longer than 2.9 kb could initiate and/or terminate in flanking sequences. They have also tested RNA from an M strain, lacking P sequences, into which a complete P factor had been introduced by injection and transposition. This strain contained a much simpler pattern of P factor RNAs, and Karess and Rubin could detect only a 2.5 kb band and a less abundant RNA of about 3 kb. These transcripts both start at about nucleotide 87 and the larger RNA seems to contain sequences at its 3′ end which are not in the 2.5 kb species. Neither of these RNAs has been associated with any function.

Fig. 1.7 Hybridization of P and I factor sequences to genomic DNA from different strains of *D. melanogaster*. (a) Genomic DNAs of the strains listed below were digested with HindIII and fractionated on a 0.7 per cent agarose gel. The DNA was transferred to nitrocellulose and hybridized with a probe made from the plasmid pπ 25.1. This carries a complete P factor (O'Hare and Rubin 1983). The phenotype of each strain, with respect to P-M dysgenesis, is indicated above each track. The band of hybridization common to all strains does not represent P DNA. It is due to hybridization of chromosomal sequences adjacent to the P factor in pπ 25.1. Note that this is the only signal from M strain DNA. The M′ strain, track 7, does contain some P sequences. The strains are, 1, *se*2014; 2, *se*F_8; 3, HJ330; 4, HJ325; 5, B2′; 6, Luminy; 7, Beanne; 8, Canton S. (b) The same filter as in (a), hybridized with a probe made from the plasmid pI771 containing 2.9 kb of DNA from the right-hand end of an I factor (Bucheton *et al.* 1984). The phenotype of each strain, with respect to I-R hybrid dysgenesis, is indicated above each track. Note that there are I factor sequences in the DNA from both I and R strains, and that the patterns of hybridization shown by different R strains are very similar. This suggests that these strains contain similar sets of I elements. Many of these elements appear to be present in each of the I strains as well. (c) Genomic DNAs from strains HJ325, 4, and B2′, 5, were digested with HindIII and PstI, and fractionated on a 0.7 per cent agarose gel. The DNA was transferred to nitrocellulose, and hybridized with a probe made from the internal 2.3 kb. HindIII.PstI fragment of the I factor (see appendix). The prominent band of hybridization seen in I strain DNA comigrates with this internal fragment. It is present in 10–15 copies per haploid genome in this strain, but only 0–1 copies in the R strain. This supports the idea that complete I factors are only present in I strains. These hybridizations were carried out by H. Sang using the experimental conditions described by Bucheton *et al.* (1984).

The mechanism of P factor transposition is not known. One of the questions to be answered is whether transposition is replicative or whether it involves excision and reinsertion. These are not necessarily mutually exclusive since one could imagine replication taking place after excision and before insertion.

Many casual observations (Bregliano and Kidwell 1983; Kidwell 1983) suggest that during dysgenesis the overall increase in P-containing sites is greater than their loss. This has been confirmed by the careful measurements of Benz (quoted in Engels and Preston 1984). This suggests that transposition is replicative, and is only compatible with non-replicative transposition if chromosomes from which P elements excise are frequently lost. Unfortunately this is difficult to exclude unequivocally. P-M dysgenesis does stimulate excision of P elements, whether or not this is part of transposition, as well as the production of gross chromosomal rearrangements. These excisions are not always precise (Voelker *et al.* 1984) and may result in loss of internal P sequences. O'Hare and Rubin (1983) and O'Hare (1985) have suggested that these internal deletions result from slippage of the replication fork during DNA synthesis associated with replicative transposition.

Engels and Preston (1984) have made a detailed study of several hundred chromosome rearrangements stimulated by P-M dysgenesis. Most of the break-points occurred at, or close to, P elements which were often lost during the rearrangements. The frequencies with which different complex rearrangements were recovered could best be accounted for by random rejoining of ends created by several chromosome breaks occurring simultaneously, rather than by a series of events each involving two breaks. Roiha *et al.* (quoted in O'Hare 1985) have studied one P-M induced inversion in detail. This occurred between two P elements and appears to have resulted from chromosome breaks at one end of each element. The sequences flanking both ends of each element are entirely conserved, but one end of one of the P elements has been lost. These results suggest that dysgenesis stimulates chromosome breaks at the ends of P elements. Precise excision could result from breaks taking place simultaneously at both ends of the same element (Engels and Preston 1984).

C. I-R DYSGENESIS AND I ELEMENTS

The transposable elements controlling I-R hybrid dysgenesis are quite distinct from P elements. Bucheton *et al.* (1984) and Sang *et al.* (1984) have analysed molecular lesions associated with eight *white* gene mutations induced by I-R dysgenesis. Two determine a bleached-white phenotype and have been found to be deletions. The remaining mutations determine a coloured eye phenotype and are associated with insertions of apparently identical 5.4 kb elements. These lie within introns or the 3′ untranslated region of the gene, and presumably allow a reduced level of *white* gene

expression. There is good reason to believe that the insertion associated with at least one of these mutations, w^{IR1}, is of an active I factor. It is very tightly linked to I factor activity, and the two have never been separated by recombination (Pelisson 1981).

Bucheton *et al.* (1984) have cloned the entire I factors associated with two of these mutations, including w^{IR1}, and have used them to investigate the distribution of these sequences in the genomes of I and R strains. All R strains so far tested contain I factor sequences, and in Southern transfer experiments DNAs from unrelated R strains show remarkably similar patterns of hybridization to I factor probes (Fig. 1.7b). These sequences must represent non-functional, defective I elements. The pattern of hybridization shown by DNA from I strains contains most of the bands common to different R strains, plus additional bands, many of which differ from one I strain to another (Fig. 1.7b). The DNA from I strains can be easily distinguished from that of R strains by digests probed to reveal large internal fragments of the I factor. These can only be seen in DNA from I strains (Fig. 1.7c). The intensity of hybridization of these internal fragments suggests that there are about 10 to 15 complete I factors per genome.

I factor probes hybridize to the chromocentric regions of polytene chromosomes from R strains, and to the chromocentre and about 15 sites on the arms of chromosomes from I strains. The positions of these euchromatic sites differ from strain to strain, suggesting that they represent transposable sequences (Bucheton *et al.* 1984; Pelisson personal communication). When taken together, these results indicate that the genomes of I and R strains contain similar arrays of non-functional I elements located near centromeres, and that I strains contain, in addition, functional and transposable I elements at sites on chromosome arms. This suggests that I elements, in contrast to P elements, are very old components of the *D. melanogaster* genome.

Vaury, Crozatier, and Bucheton (personal communication) have studied incomplete I elements from both I and R strains. The relationship between complete and incomplete I elements is more complex than that between complete and incomplete P elements. Some I elements have lost one or other end of the I factor, or some internal sequences. One has an internal region substituted by unrelated DNA and some have more than one rearrangement.

Fawcett *et al.* (1986) have determined the complete base sequence of the I factor from w^{IR1}. It is flanked by a target site duplication, but does not have terminal repeats (see appendix). They have sequenced the ends of the five other I factor insertions within the *white* gene, and this pattern is common to all of them. The length of the target site duplication varies from one insertion to another and ranges from 12 to 14 bp (see appendix). There is a hot-spot for I factor insertion within the *white* gene, and three of the insertions are at exactly the same position. This differs from the hot-spot for P element insertion. The ends of these I factors are well conserved. At the 3′ end of the putative coding strand there is a short run of TAA triplets, the length of which

varies from one I factor to another. Runs of four to seven copies have been found.

All of the coding information of the I factor is within one strand which has two large ORFs. The first is 1278 bp long and has an ATG as the fourth codon. The second is 3258 bp long and has ATGs as the second and fourth codons.

Transposition of I factors is probably replicative since they increase in number during transposition (Bregliano and Kidwell 1983), but again it is impossible to rule out entirely the possibility of excision followed by chromosome loss. There is no evidence that I factors can excise precisely during hybrid dysgenesis, although they are unstable under these conditions. Pelisson (1981) has found that the w^{IR1} mutation does not revert in dysgenic individuals, but can give rise to other mutant alleles. Some of these have deletions a few kilobases long. They start at one end of the I factor and remove adjacent *white* locus sequences (Lynch, Walsh, and Kellet personal communication).

The structure of I factors resembles that of mammalian LINES and F elements on *D. melanogaster*, since each lacks terminal repeats and has an A-rich sequence at the 3′ end of one strand. This suggests that all these elements may transpose by similar mechanisms. It is unlikely that the A-rich sequence of I factors is degraded polyA as the repeating TAA motif is conserved absolutely, and is not preceded by an appropriate polyadenylation signal. I factors are known to determine a function required for their own transposition, and this information is presumably contained within one, or both, of the ORFs mentioned above. The second of these encodes a polypeptide with homology to viral reverse transcriptases, reinforcing the idea that I factor transposition involves an RNA intermediate (Fawcett *et al.* 1986).

VI. Hobo—a transposable element with short inverted repeats

The first *hobo* element, *hobo101*, was found as a 1.3 kb insertion 33 bp upstream of the start of an *sgs*-4 allele segregating in the Stromsvereten 8 strain of *D. melanogaster* (McGinnis *et al.* 1983). The results of Southern transfer experiments, in which genomic DNAs of several strains were digested with XhoI to reveal internal fragments of *hobo* elements, indicated that there are 20–50 copies of this element per haploid genome. The sizes and number of elements varied from strain to strain, except that DNA from all strains gave a prominent band of hybridization corresponding to *hobo* elements of about 3 kb.

Streck, MacGaffey, and Beckendorf (personal communication) suggest that smaller *hobo* elements, such as *hobo101*, may be related to these 3 kb elements by internal deletions/insertions, just as incomplete P factors are related to complete P factors. They have cloned one 3 kb element, *hobo108*, and have

determined its complete base sequence, as well as that of *hobo101* (McGinnis *et al.* 1983). Both elements have terminal inverted repeats 12 bases long, and are flanked by 8 base direct repeats. This direct repeat is known to be a duplication of the target site in the case of *hobo101*. The sequence of *hobo108* contains one long ORF extending 1932 bp from the first ATG. This is preceded by a potential promoter region containing sequences closely related to the consensus TATA and CAAT boxes characteristic of polymerase II promoters. Streck *et al.* (personal communication) suggest that this ORF may code for a transposase.

The sequence of *hobo101* is very similar to that of *hobo108* except for a single deletion/insertion of 1.7 kb including two-thirds of the long ORF. The ends of three other short *hobo* elements have been sequenced and are identical to those of *hobo101* and *hobo108*. These elements are also flanked by 8 bp direct repeats. There is no obvious homology between these putative target site duplications except that in four out of five cases the last two bases are AC. This sequence occurs as base 7 and 8 of the terminal inverted repeats of these elements, but there is no evidence to suggest that this is anything but fortuitous.

McGinnis *et al.* (1983) have isolated a derivative, S8D, of the Stromsvereten 8 strain, which is homozygous for the *sgs*4 allele associated with *hobo101*. Expression of *sgs*4 is reduced 50–100 fold in this stock but its time of expression in development is unaltered. This reduction in *sgs*4 expression probably reflects the fact that *hobo101* has inserted between the TATA box of the *sgs*4 promoter and any upstream regulatory sequences. This may reduce the efficiency of either RNA polymerase binding or initiation of transcription. McGinnis *et al.* (1983) have found two size classes of *sgs*4 transcripts which seem to initiate within *hobo101* in S8D, and are associated with DNAseI hypersensitive sites. These probably result from fortuitous initiation events since the truncated long ORF in *hobo101* is in the opposite orientation to that of *sgs*4. They could detect no other *hobo* transcripts.

VII. Unanswered questions

Many questions remain to be answered concerning transposable elements in *D. melanogaster*, despite the large amount of descriptive information already available. Except for *copia*-like elements, very little is known about mechanisms of transposition. This can now be investigated directly by introducing mutant elements into flies by transformation. A start has been made with P elements, and no doubt others will follow shortly.

Nothing is known, at the molecular level, about factors which control transposition frequencies. Transposition of P and I elements is increased in

the progeny of appropriate dysgenic crosses, but the molecular basis of this is unknown. Gerasimova and her colleagues have found strains of flies in which several different transposable elements are unstable, suggesting that transposition may occur in bursts (Gerasimova 1983; Gerasimova *et al.* 1984). This could explain the apparent paradox that, within any particular strain, the rate of transposition of most elements is very low (Tchurikov *et al.* 1981; Young and Schwartz 1981), and yet between strains the chromosomal distribution of elements is different.

Little is known about the origin of transposable elements, or about the role they play in genome evolution. *Copia*-like elements are clearly related to retroviruses, but has one evolved from the other, or do they have similar evolutionary histories without any direct connection? From what sequences do transposable elements evolve? Some elements, such as the *copia*-like and I elements, appear to be very old components of the genome since they are found in all strains of *D. melanogaster* and in its sibling species (Martin *et al.* 1983; Bucheton, personal communication) while others, such as P elements, may have entered *D. melanogaster* recently, and have originated in a distantly related species (Daniels and Strausbough, 1985).

Acknowledgements

We are grateful to all our colleagues who sent us information about various transposable elements, and to M. Lynch, M. Pritchard and J. Prosser who read and criticized the manuscript. We are also grateful to H. Sang for providing the data for Fig. 1.7. The figures were drawn by A. Wilson, and were photographed by G. Brown. Some of the experiments described in this review were carried out with the support of Project Grants from the Medical Research Council. D. H. F. holds a Research Studentship from the Science and Engineering Research Council.

Appendix

Detailed information on all the transposable elements listed in Table 1.1. is given on the following pages. The elements are described in order according to size. Restriction maps are given for each element and, where possible, the base sequences of their termini. We have tried to orient the maps of *copia*-like elements so that transcription of RNAs analogous to retroviral genomic RNA would be from left to right. There is insufficient information to do this for the following elements: *1731*, *NEB*, *3S18*, *BEL* and *springer*. The LTRs of *copia*-like elements are indicated by boxes. If two restriction sites are shown

above a T symbol, then their order is not known. We have indicated which enzymes are known not to cut each element.

Sancho 2

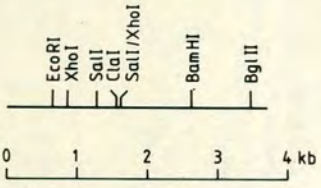
Length: 2.6 kb (1).

Map: From the map published by Campuzano *et al.* (1). There are no sites for the enzymes BglII and SalI.

Approximate copy number: 30 (1).

Comments: First described as an insertion on a chromosome carrying the Hw^1 mutation, although it probably does not contribute to the mutant phenotype. Only one copy of the element has been cloned. The 1.7 kb EcoRI–HindIII fragment is repeated about 30 times in the genomes of strains Oregon R and Vallecas. It cross-hybridizes with *Sancho 1* (1).

Reference: (1) Campuzano *et al.* (1985).



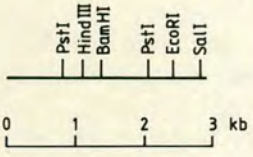
Jockey

Length: 2.8 kb (1).

Map: From the map published by Mizrokhi *et al.* (1) and Georgiev (2). The SalI site is only present in a few *Jockey* elements (1). There are no sites for the enzymes BglII, KpnI and XhoI (2).

Comments: Only one *Jockey* element has been described in detail (1). It was found inserted within the *mdg*4/*gypsy* element associated with the $ct^{MR2pN10}$ mutation. This was derived from ct^{MR2}. *Jockey* elements are moderately repeated within the genome, and have different chromosomal locations in different strains. The structure of *Jockey* elements themselves is well conserved. No homology has been detected between the ends of the element which has been cloned.

References: (1) Mizrokhi *et al.* (1985).
(2) Georgiev (personal communication).



P

Length: 2907 bp (1).

Map: From the sequence of a complete P factor published by O'Hare *et al.* There are no sites for the enzymes BamHI, HpaI and SmaI.

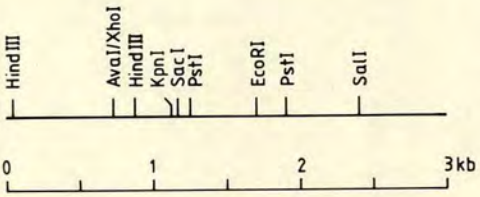
Terminal inverted repeat: 31 bp (1).

Target site duplication: 8 bp (1).

Approximate copy number: 0–50 (2).

Comments: First described by Rubin *et al.* (3) as being associated with *white* gene mutations induced by P-M hybrid dysgenesis. This sequence was reported by O'Hare *et al.* (1). This map is of a complete P factor; many shorter P elements have been described. These are mostly related to the complete P factor by internal deletions (1). P elements have been found associated with P-M induced mutations at many loci. An 8 bp consensus target sequence, GGCCAGAC, has been proposed by O'Hare *et al.* (1).

References: (1) O'Hare *et al.* (1983).
(2) Bingham *et al.* (1982).
(3) Rubin *et al.* (1982).



Ends of a P factor.

Left-hand end

```
1    CATGATGAAATAACATAAGGTGGTCCCGTCGAAAGCCGAAGCTTACCGAA   50
```

Right-hand end

```
2858  TTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATCATG  2907
```

hobo

Length: 3016 bp (1).

Map: From the sequence determined by Streck *et al.* (1). There are no sites for the enzymes BamHI, HpaI, KpnI, SacI and SmaI.

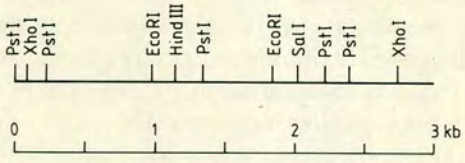
Terminal inverted repeats: 12 bp (1, 2).

Target site duplication: 8 bp (1, 2).

Approximate copy number: 20–50 (2).

Comments: First described by McGinnis *et al.* (2) as being associated with an *sgs*4 gene in the strain Stromsvereten 8. The sequences given above are from *hobo108* (1). This orients the whole element so that the long ORF reads from right to left.

References: (1) Streck, R. D., MacGaffey, J. E. and Beckendorf, S. K. (personal communication).
(2) McGinnis *et al.* (1983).



Ends of a hobo element.

Left-hand end

```
1   CAGAGAACTGCAGCCCGCCACTCGCACTCTACGTCCACCCGATAAACACT   50
```

Right-hand end

```
1224  TGTAGGGTGTGAGTCGAGTGGTAAAAAAGTGCCACCCTTGCAGTTCTCTG   1273
```

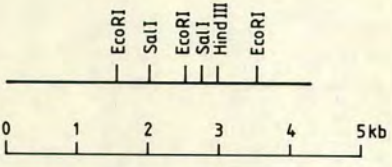
Doc

Length: 4.3 kb (1).

Map: From the map published by Bender *et al.* (1).

Comments: Only one *Doc* element has been described. It was found inserted in the Bithorax region on a chromosome carrying the bx^3 mutation, but not in the corresponding position on other chromosomes. It is not responsible for the bx^3 mutation, which is associated with a *gypsy* insertion, since bx^3 and *Doc* have been separated by recombination (1).

Reference: (1) Bender *et al.* (1983a).



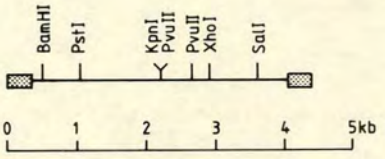
1731

Length: 4.4 kb (1).

Map: This map was determined by F. Peronnet (1). There are no sites for the enzymes EcoRI, HindIII and XbaI.

Comments: The first *1731* elements were identified because their transcription in tissue culture cells decreases after the addition of ecdysone (2). Heteroduplex experiments between fragments containing the ends of a *1731* element indicate that it has LTRs about 350 bp long (1, 2). The number of copies of *1731* elements is higher in tissue culture cells than in whole flies. In tissue culture cells they are transcribed into full-length RNAs, and are complementary to extrachromosomal circular DNAs (2). For these reasons they are regarded as being *copia*-like elements (2).

References: (1) Peronnet (personal communication).
(2) Peronnet *et al.* (1986).



Kermit

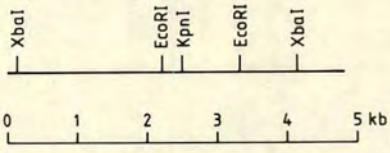
Length: 4.8 kb (1).

Map: From a map provided by Bender (2). There may be additional EcoRI sites at the ends of the element. There are no sites for the enzymes BamHI, HindIII and SalI.

Approximate copy number: 30 (1).

Comments: Only one copy of this element has been described. This was found within the 87E1-6 region in Canton S DNA, but not in Oregon R DNA (1).

References: (1) Bender *et al.* (1983b).
(2) Bender, W. (personal communication).



Sancho 1

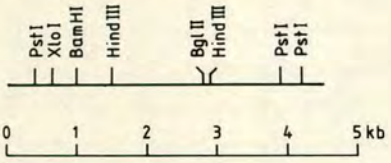
Length: 4.5 kb (1).

Map: From the map published by Campuzano *et al.* (1). There are no sites for the enzymes EcoRI and XhoI.

Approximate copy number: 50 (1).

Comments: First described as an insertion in a chromosome carrying the sc^{D1} mutation, although it probably does not contribute to the mutant phenotype. Only one copy of the element has been cloned. The 1.45 kb HindIII fragment is repeated about 50 times in the genomes of strains Oregon R, Canton S and Vallecas. It cross-hybridizes weakly with *gypsy* and with *Sancho 2* (1).

Reference: (1) Campuzano *et al.* (1985).



copia

Length: 5146 bp (1, 11).

Map: From the sequence of a *copia* element published by Mount and Rubin (1). There are no sites for the enzymes AvaI, BamHI, BglII, KpnI, SacI, SalI, SmaI and XhoI.

Terminal inverted repeat: 13/17 bp (2).

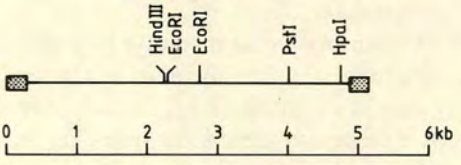
Target site duplication: 5 bp (3).

Approximate copy number: 60 (4, 5).

Comments: First described by Finnegan *et al.* (4) as a sequence complementary to abundant polyA$^+$ RNA in tissue culture cells. This sequence was reported by Levis *et al.* (2). The region containing the initiation sites for *copia* transcription are underlined (10, 11). The most probable major initiation sites are marked by '*' above the bases concerned (10, 11). Emori *et al.* (11) have suggested that the boxed sequence contains the polyadenylation signal. *Copia* insertions are associated with the following mutations, w^a (6, 7), $w^{hd81b11}$ and $w^{hd81b25}$ (8), and Bx^{46} (9).

References: (1) Mount and Rubin (1985).
(2) Levis *et al.* (1980).
(3) Dunsmuir *et al.* (1980).
(4) Finnegan *et al.* (1978).
(5) Potter *et al.* (1979).
(6) Gehring and Paro (1980).
(7) Bingham and Judd (1981).

(8) Rubin *et al.* (1982).
(9) Mattox and Davidson (1984).
(10) Flavell *et al.* (1981).
(11) Emori *et al.* (1985).



LTR of a copia element.

```
   1  TGTTGGAATATACTATTCAACCTACAAAAGTAACGTTAAACAACACTACT   50
  51  TTATATTTGATATGAATGGCCACACCTTTTATGCCATAAAACATATTGTA  100
 101  AGAGAATACCACTCTTTTTATTCCTTCTTTCCTTCTTGTACGTTTTTTGC  150
 151  TGTAAGTAGGTCGTGGTGCTGGTGTTGCAGTTGAAATAACTTAAAATATA  200
 201  AATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAA  250
 251  GGAAATATAAATTATAAATTACAACA  276
```

I

Length: 5371 bp (1).

Map: From the sequence reported by Fawcett *et al.* (1). There are no sites for the enzymes BamHI, EcoRI, SacI, SalI, SmaI and XhoI.

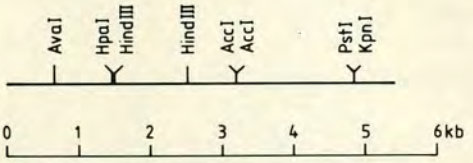
Terminal inverted repeat: None.

Target site duplication: Variable, 12–14 bp duplications have been observed (1).

Approximate copy number: 0–15 complete I factors, plus 30 I elements (2).

Comments: First described by Bucheton *et al.* (2) as insertions associated with *white* gene mutations induced by I-R hybrid dysgenesis. This sequence was reported by Fawcett *et al.* (1). The number of copies of the 3′ terminal TAA is variable. I factors have been found associated with other I-R induced mutations (3), and with the spontaneous mutation bx^{F31} (4).

References: (1) Fawcett *et al.* (1986).
(2) Bucheton *et al.* (1984).
(3) Sang *et al.* (1984).
(4) Peifer, M. and Bender, W. (personal communication).



Ends of an I factor.

Left-hand end

1 CATTACCACTTCAACCTCCGAAGAGATAAGTCGTGCCTCTCAGTCTAAAG 50

Right-hand end

5323 AGTTAGTCTAGTTTTGTAAACTATTCTATCTATCATAATAATAATAATAA 5372

*mdg*3

Length: 5.4 kb (1).

Map: The reverse of the map published by Ilyin *et al.* (1). There are no sites for the enzyme BamHI.

Terminal inverted repeat: 4/5 bp (2, 3).

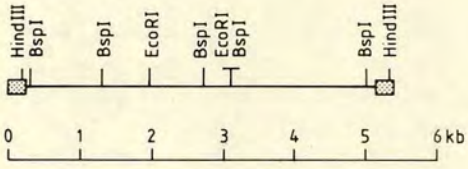
Target site duplication: 4 bp (2, 3).

Approximate copy number: 15 (1).

Comments: First described by Ilyin *et al.* (1, 4) as being complementary to double-stranded RNA from tissue culture cells. This sequence is the reverse complement of that published by Bayev *et al.* (5). This orients the complete element so that the most likely primer binding site is next to the left-hand LTR and the putative purine-rich sequence is next to the right-hand LTR (6). The reported direction of major transcription would, however, be right to left (1). Bayev *et al.* (5) reported the sequence of the LTRs of one *mdg*3 element. They interpreted their data as indicating that 5 bp of the target site DNA had been duplicated and that the LTRs were 268 bp long, with 18 bp inverted repeat sequences at the ends of the whole element, but not at both ends of each LTR. This could not be confirmed because the corresponding target site had not been cloned. Mossie *et al.* (2) have cloned a free *mdg*3 LTR, plus a complete element together with its empty target site. They interpret all the available sequence information as indicating a 267 bp LTR with 4/5 bp terminal inverted repeat and a 4 bp target site duplication. Bayev *et al.* (3) have cloned

and sequenced the ends of four more copies of *mdg*3. Their data support these conclusions.

References: (1) Ilyin *et al.* (1980*a*).
(2) Mossie *et al.* (1985).
(3) Bayev *et al.* (personal communication).
(4) Ilyin *et al.* (1980*d*).
(5) Bayev *et al.* (1980).
(6) Kugimya *et al.* (1983).



LTR of an mdg3 element.

```
  1  TGTAGTAGGCTGCTCCTTCTACCCTCTTCCTTTACTCTTAGTCATACATA   50
 51  CCTAATTATACATAGCCAATCTAGTCATAAGCTTATACACTCATACACCC  100
101  ATCCTTAACATACAAATATTATCGAGAAACTTATCGACTAATCGACTCGC  150
151  CACTCTGCAGAGAGCGCGGCAGTCAGTCGCTGTTGAACCAAGCTAAAGGA  200
201  CAGATCAAAAATAAAAGAGACACGTGAAATTGTATTAGAATATTAACTTC  250
251  TGTAAACGGCGGCTAAA  267
```

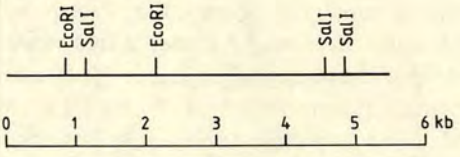
NEB

Length: 5.5 kb (1).

Map: From the map published by Paro *et al.* (1). There are no sites for the enzymes BamHI and XhoI.

Comments: Only one copy of the *NEB* element has been described. It was found on the large transposable element TE98 near the end carrying the *roughest* gene. The ends of this element cross-hybridize, but no inverted repeats could be detected by electron microscopy. This suggests that *NEB* is a *copia*-like element. *NEB* elements occur at multiple dispersed sites in the genome and are located at different positions in different strains. An incomplete *NEB* element has been found near one end of the large transposable element TE77 (1).

Reference: (1) Paro *et al.* (1983).



3S18

Length: 6.5 kb (1).

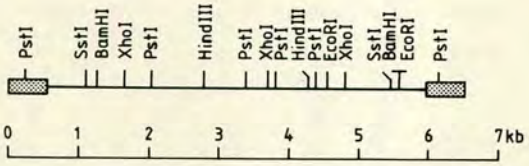
Map: From the maps published by Bell *et al.* (1) and Mattox and Davidson (2). There are no sites for the enzymes BglII, SmaI and XbaI.

Terminal inverted repeats: 5 bp (3).

Approximate copy number: 15 (1).

Comments: First identified as an insertion within the non-transcribed spacer of an rDNA repeat (1, 4). Restriction and Southern hybridization data suggest that this element is flanked by direct repeats about 500 bp long (1). A *3S18* probe hybridized to 12 euchromatic sites and the chromocentre of salivary gland chromosomes from larvae from a cross between the gt^1 and gt^{X11} strains. The distribution of sites was different in these two strains (1). These data suggest that *3S18* is a *copia*-like element. *3S18* elements have been found associated with the Bx^J (2) and w^{zm} (3) mutations. Little, if any, RNA complementary to an internal fragment of *3S18* could be found in embryos, larvae, and adults, or in Schneider line 2 tissue culture cells (1).

References: (1) Bell *et al.* (1985).
(2) Mattox and Davidson (1984).
(3) O'Hare *et al.* (1984).
(4) Fabijanski and Pellegrini (1982).



297

Length: 6995 bp (5).

Map: From the sequence of a *297* element determined by Saigo *et al.* (5). There are no sites for the enzymes BamHI, PstI, SacI, SalI, SmaI and XbaI.

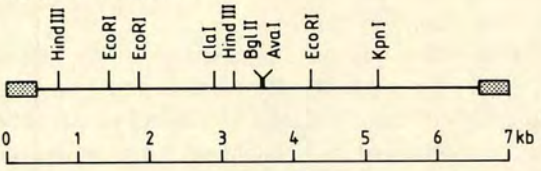
Terminal inverted repeat: 4/5 bp (1).

Target site duplication: 4 bp (1, 2).

Approximate copy number: 30 (3).

Comments: *297* elements were first described by Potter *et al.* (3) but had been identified by Wensink and Rubin (personal communication) as complementary to an abundant polyA$^+$ RNA in tissue culture cells. This sequence was reported by Ikenaga and Saigo (1). Insertions are preferentially at the sequence ATAT (2). It is closely related to *17.6* elements. The sequences of the LTRs of *297* and *17.6* elements are similar, and heteroduplexes between *297* and *17.6* elements show a region of 1.7 kb of homology at the right-hand ends of each (4).

References: (1) Ikenaga and Saigo (1982).
(2) Spradling and Rubin (1981).
(3) Potter *et al.* (1979).
(4) Kugimya *et al.* (1983).
(5) Saigo (personal communication).

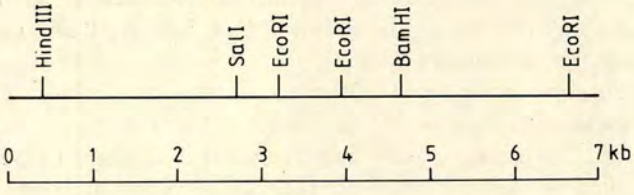


LTR of a 297 element.

```
  1  AGTGACGTATTTGGGTGGACCAAACCAGCCACTTCCATTATTTCAAAGAA   50
 51  ATCAGTAATGCACTCTAGTAATTTTCCATAACTGTATCCCAGCTGCGCAG  100
101  ACTCGTTTACCTTTTGCAGCGCAGCGTTCTTTGTAAACATCCTAAAGACC  150
151  TGCCTAAGCAGATTTGACTGCCCTCTTTCAACGCTACCTAATCCTAAGAA  200
201  CCCAAGAGCGAGGCTCTCCCGAAATACAAATATTGTTCAAATACTGAGGC  250
251  TTCTCCTCAATCCAATTTGCATTTGATTTTTAGTCTTAAGCTGAGATCCA  300
301  AAGAATAAAGTCGTGAAACTATTTCTCCTAAAAACTATTTTTTATTTCTT  350
351  GGCGTTGTCCTTAGTCAACTGACGGGACATTAGTTCAGACTCATACATAA  400
401  AACAACAATTTTACT  415
```

Delta 88

Length: 7 kb (1).
Map: From the map published by Karch *et al.* (1).
Comments: First described by Karch *et al.* (1) as an insertion associated with the *tuh*-3 mutation. This is a moderately repetitive element.
Reference: (1) Karch *et al.* (1985).

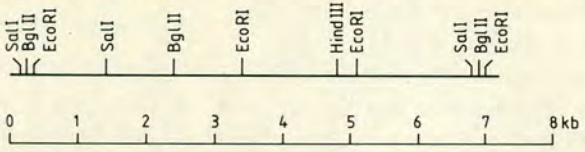


Calypso

Length: 7.2 kb (1).
Map: From the map supplied by Bender and Curtis (1).
Approximate copy number: 10–20 (1).
Comments: First described as an insertion associated with the ry^{301} muta-

tion (1). Three other copies have been cloned. They are identical to the ry^{301} element as judged by heteroduplex analysis, but show some small variations in their restriction maps (1).

Reference: (1) Bender, W. and Curtis, D. (personal communication).



Harvey

Length: 7.2 kb (1).

Map: From the map supplied by Bender and Peifer (1).

Comments: First described by Bender and Peifer (1) as being associated with the bx^{8} mutation. The results of whole genome Southern experiments indicate that this element is repeated in the genome, and that the internal SalI–SalI fragment is conserved in length. There is some homology between the ends of the element.

Reference: (1) Bender, W. and Peifer, M. (personal communication).

BEL

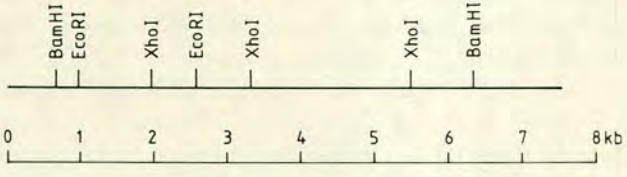
Length:7.3 kb (1).

Map: From the map published by Goldberg *et al.* (1).

Approximate copy number: 25 (1).

Comments: A *BEL* element has been found associated with the w^{a4} mutation (1). *BEL* elements are located at about 25 sites throughout the *D. melanogaster* genome, and their distribution differs from strain to strain (1). On the basis of these data, and the fact that the ends of a *BEL* element cross-hybridize, Goldberg *et al.* (1) have suggested that these are *copia*-like elements. It is not clear whether the terminal repeats are direct or inverted.

Reference: (1) Goldberg *et al.* (1983).



HMS Beagle

Length: 7.3 kb (1).

Map: The reverse of the map published by Snyder *et al.* (1) There are no sites for the enzymes AvaI, BamHI, HindIII and KpnI.

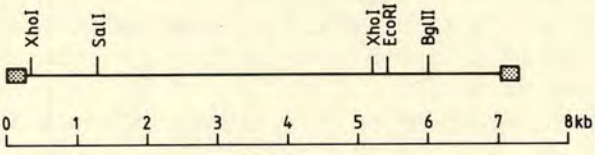
Terminal inverted repeat: 6/7 bp (1).

Target site duplication: 4 bp (1).

Approximate copy number: 50 (1).

Comments: First described by Snyder *et al.* (1) as an insertion within the promoter region of the cuticle protein gene *CP3* (1). This insertion is associated with loss of *CP3* activity. The sequence above is the reverse complement of that published by Snyder *et al.* (1). This puts a possible primer binding site next to the left-hand LTR and a putative purine-rich sequence next to the right-hand LTR. Only one copy of this element has been described.

Reference: (1) Snyder *et al.* (1982).



LTR of an H. M. S. Beagle element.

```
  1  AGTTATTGCCCTGCAATTGATTCTCTAACATCTTGTGGTTCCACATAGTC  50
 51  TCCGCTGCCATCAACGCCAACGAACGGTTAAGCGCGACATCGACACTTCT 100
101  GCGCTGCGCCGCGGCCGACGCCTGCTGCGCCACTGCCGACGACTTCACTT 150
151  GATTGCTAGGGACTTAGGGAAACATTTTGTACGCTAGATTTAGTTTCGAA 200
201  TGATAAATTGCAATAAACGGTCGCTTGCGATCTTCAAAATCAAATCGATA 250
251  ACTGTAATTATTAACT 266
```

*mdg*1

Length: 7.3 kb (1).

Map: The reverse of the map published by Ilyin *et al.* (1).

Terminal inverted repeat: 13/16 bp (2).

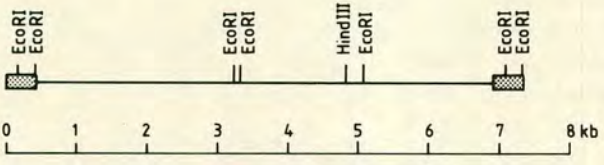
Target site duplication: 4 bp (2).

Approximate copy number: 25 (1).

Comments: First described by Ilyin *et al.* (3, 4) as being complementary to abundant polyA$^+$ RNAs. The sequence given above is the reverse complement of that published by Kulguskin *et al.* (2). This puts a possible primer binding site adjacent to the left-hand LTR, and a purine-rich sequence adjacent to the right-hand LTR. The direction of major transcription is left to

right (5). Fourteen of the 18 bases of the putative primer binding sites of *mdg*1 and *412* elements are identical. The 27 bases immediately adjacent to the left-hand LTRs of these elements are also identical (6).

References: (1) Ilyin *et al.* (1980*b*).
(2) Kulguskin *et al.* (1981).
(3) Ilyin *et al.* (1978).
(4) Georgiev *et al.* (1978).
(5) Ilyin *et al.* (1980*c*).
(6) Will *et al.* (1981).



LTR of an mdg1 element.

```
  1  TGTAGTTAATTTGAATTCTAATACTTCTGATGTAGTTAATAAAGTCTCAA  50
 51  AACGCAGTTGGTCATTTTATCTTTATTTGGTTTTTATTAAGCGAGTGTAC 100
101  ATTCCTGATCTATTCCTGATTTATAACTGATCTTAGTGTGGTCTACAGGG 150
151  AGATCTCTTGTGACAGCCAAGTGCTGACACTAGCAAATTCTGCAATATGT 200
201  ATGTATGAAGTTCATACTCGATAACCAATGGGAGTCGAGTCCGACCCCTA 250
251  AAGGCGTACATCCTGAATTCGCATATTTAGTATTAGGGTGTATCTAAAGA 300
301  TCTACTAGGGTGACCCTAAGGAATTAGGGTGGTCCTAAGTTTACTTATTA 350
351  GTTGACTTATTATTATGATTCATATTATAATTATTATTAATATTATTATT 400
401  ATTATTGTTATTATTATTGTTATTATATTCGTATATACTACA 442
```

*mdg*4/*gypsy*

Length: 7.3 kb (1, 2).

Map: From maps published by Bayev *et al.* (1) and Mattox and Davidson (2). There are no sites for the enzymes BamHI and SalI.

Terminal inverted repeat: 4/5 bp (1, 3).

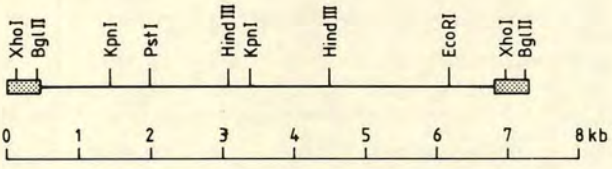
Target site duplication: 4 bp (1, 3).

Approximate copy number: 10 (4, 8).

Comments: Described by Ilyin *et al.* (6) as *mdg*4, a sequence complementary to double-stranded RNA from tissue culture cells, and by Bender *et al.* (5) as *gypsy*, an insertion associated with the mutations bx^3, bx^{34e}, bxd^1, bxd^{55i} and bxd^{51j}. *Gypsy* insertions are also associated with the mutations ct^6 (4), f^1 (7), Bx^2 (2), y^2 (9) and Hw^1 and Hw^{BS} (10). Mutations sc^1, sc^{D1}, sc^{D2}, sc^{L3} and sc^{3B} are associated with *gypsy* elements inserted at apparently the same site and in the same orientation (9). Modelell *et al.* (4) have shown

by *in situ* hybridization that *gypsy* insertions are associated with many mutations suppressed by *su(Hw)*. This sequence was reported by Bayev *et al.* (1). Freund and Meselson (3) have reported an equivalent sequence of 482 bp.

References: (1) Bayev *et al.* (1984).
(2) Mattox and Davidson (1984).
(3) Freund and Meselson (1984).
(4) Modelell *et al.* (1983).
(5) Bender *et al.* (1983*a*).
(6) Ilyin *et al.* (1980*d*).
(7) Parkhurst and Corces (1985).
(8) Tchurikov *et al.* (1981).
(9) Campuzano *et al.* (1986).
(10) Campuzano *et al.* (1986).



LTR of an mdg4/gypsy element.

```
   1  AGTTAACAACTAACAATGTATTGCTTCGTAGCAACTAAGTAGCTTTGTAT   50
  51  GAACAATGCTGACGCGCCAGAATTGGGTTCAACGCTCCACGCGAAGAATG  100
 101  CCTGGCAGCGGAAAGCTGACACTTCCTACCGGGAGTGTTGCTTCACGCTG  150
 151  CAAGAAATGCTGGCGGCTGCCGACTTGTGGCGGCGCGATGCATTGCTCGA  200
 201  GGGTAAACTTAGTTTTCAATATTGTCTTCTACTCAGTTCAAATCTTGTGT  250
 251  CGAAATAAACCACAGCTTGCTCCGGCTCATTGCCGTTAAACATCATTGTT  300
 301  CTTATTTACAATCAAATCGCTATCGCCACAAGGCTAGTGATAATAACTAA  350
 351  GGGGGCGAAGTCAAGCCCTCCAACCTAATCTCCATAAACAGTGTCTAAGA  400
 401  CGAACCTCAGCGAAAGAAGGAAGATCTCTAGACCTACTGGAAATAACATA  450
 451  ACTCTGGACCTATTGGAACTTATATAATT  479
```

17.6

Length: 7439 bp (1).

Map: From the sequence of a *17.6* element published by Saigo *et al.* (1). There are no sites for the enzymes AvaI, KpnI, SacI, SmaI and XhoI.

Terminal inverted repeat: none (2).

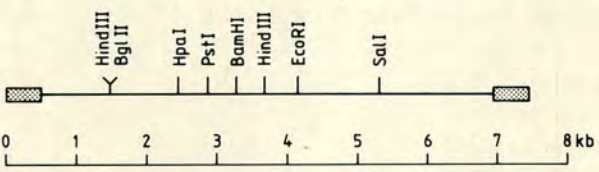
Target site duplication: 4 bp (2, 3).

Approximate copy number: 40 (2).

Comments: First described by Saigo *et al.* (4) as a sequence inserted into histone genes and hybridizing to *297* elements. This sequence was reported by Kugimya *et al.* (2). The sequences of the LTRs of *17.6* and *297* elements are similar. Heteroduplexes between *17.6* and *297* elements show a 1.7 kb region of homology between the right-hand ends of each (2). All insertions studied so far are associated with duplications of the target site sequence ATAT. As a result *17.6* insertions are preferentially within the consensus promoter sequence TATATA (3).

References: (1) Saigo *et al.* (1984).

(2) Kugimya *et al.* (1983).
(3) Inouye *et al.* (1984).
(4) Saigo *et al.* (1981).



LTR of a 17.6 element.

```
  1  AGTGACATATTCACATACAAAACCACATAACATAGAGTAAACATATTGAA   50
 51  AAGCCGCATACGTAAACAATAAGTGACCACCATGCTAATGTGGATCAAAT  100
101  AACAAAAATATCCACTCTGCATTTTGACACCCCCATACTGTATGCCATCT  150
151  GCGCAGTATGCATTCTAATAAACAAATTCTTTGACAGCGGCACTTAGCCA  200
201  TTCTTGTAAACAAATCTTAAAGTCTGCCTGCTCTCTCTGAGGCTTCTCCT  250
251  CCACTTAAGAATCCAAGAGCAATGCTCTCCCAAAAACACTAACATATTCT  300
301  TTAAGCAAGCACAGAGGCTTCTCCTCATTTTCACTTTCATTTGATTTTCA  350
351  GTCTTAAGCTGAACGTTAATCAATAAACAACACAATCGATACCGAAATTT  400
401  TGATTCGTTTTATTTTGGCAAAACTCAATTTTCAGCGTTGGTCTTAGTTC  450
451  ATATTCGGAACGGTCCATTTAATAGACTCAAAACTATTTATTGCAACCAT  500
501  TTATTTGCAATT  512
```

412

Length: 7.6 kb (1).

Map: From the map published by Shepherd and Finnegan (1). There are no sites for the enzymes BamHI, SacI and SalI.

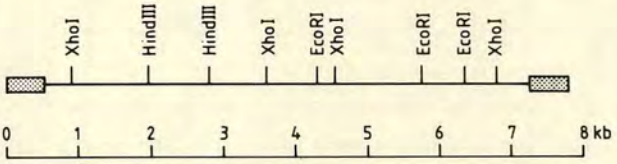
Terminal inverted repeat: 5/6 bp (2).

Target site duplication: 4 bp (2).

Approximate copy number: 40 (3, 4).

Comments: First described as being complementary to abundant polyA$^+$ RNA in tissue culture cells (6). This sequence was reported by Will *et al.* (2). Rare *412* elements have 571 bp LTRs, the first 482 bp of which correspond to the sequence shown (2). Shepherd and Finnegan (1) have suggested that *412* elements insert preferentially at the sequence NTNG. This is based on analysis of four elements only. Finnegan *et al.* (2) reported that the direction of major transcription in K_c tissue culture cells is right to left. Prosser and Finnegan (unpublished data) have found that both strands are about equally represented in polyA$^+$ RNA from adults. A *412* element is associated with the following mutations *bx*^{1} (5), v^1, v^2 and v^k (7). Fourteen of the 18 bases of the putative primer binding sites of *412* and *mdg*1 elements are identical. The 27 bases immediately adjacent to the left-hand LTRs of these elements are also identical (2).

References: (1) Shepherd and Finnegan (1984).
(2) Will *et al.* (1981).
(3) Finnegan *et al.* (1978).
(4) Potter *et al.* (1979).
(5) Bender *et al.* (1983*a*).
(6) Rubin *et al.* (1976).
(7) Searles and Voelker (1985).



LTR of a 412 element.

```
  1  TGTAGTATGTGCCTATGCAATATTAAGAACAATTAAATAAAATAGCATAT  50
 51  TAACTTATGGCAGCACTTTGTTGCTATGTTTATGTTTATGTTTATGCACG 100
101  CAGTTAGGCCAGGGCGGATGTAACATGATCACCCACTCGAAGGCAAAAAG 150
151  TATAAGTGCATGGTCAGCATTCACACGCCGACCAAATACATATTACATAC 200
201  GTACATACATATCTCGCTCTCCCGATAAGCCTAGATATATAAGATATACA 250
251  TAAGAACGCCGCTCCGCTGCTGGCGTACCCGGCAGCGCAGCTACGCGGAT 300
301  TAGCCTAAGTCCAAATATATTAAAAACTGTAAAATCGGAGAGACTCTGTA 350
351  GACGTTGAGCGGACAGAACCATTTCTGCCTACTCTAAAATCAAAAGAAGA 400
401  AATTGAATAAATATATGTCAGCCCGACGGCTGCCTTCAACTTAAAACGGA 450
451  CTTGTGTTCTGAATTGGAGTTCATCATTACA 481
```

BS

Length: 8 kb (1).

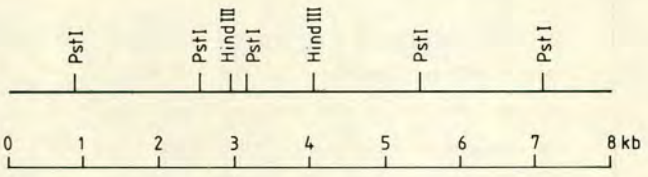
Map: From the map published by Campuzano *et al.* (1). There are no sites for the enzymes BamHI, EcoRI and SalI.

Terminal inverted repeat: About 2.5 kb for the only BS element so far described (1).

Approximate copy number: 15 (1).

Comments: First described as a sequence inserted within the *gypsy* element associated with the Hw^{BS} mutation (1). It has long inverted terminal repeats as judged by hybridization and restriction analysis. The internal 1.7 kb PstI fragment is repeated about 15 times in strain Oregon R (1).

Reference: (1) Campuzano *et al.* (1986).



B104/roo

Length: 8.7 kb (1).

Map: From the maps published by Scherer *et al.* (1) and Swaroop *et al.* (5).

Terminal inverted repeat: 3 bp (1).

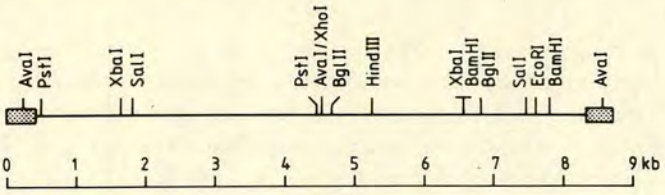
Target site duplication: 5 bp (1).

Approximate copy number: 80 (1).

Comments: Described as *B104* by Scherer *et al.* (1, 2), and as *roo* by Meyerowitz and Hogness (3). *B104* elements were found because they are complementary to abundant polyA$^+$ RNA in embryos (2), while a *roo* element was found inserted near the *sgs*3 gene (3). This sequence was reported by Scherer *et al.* (1). The boxed sequence is the probable polyadenylation signal. The base marked '*' is the last base before polyA in a *B104* cDNA (1). *B104/roo* elements have been found associated with the following mutations, w^{bf1} and w^{sp1} (4), *Gl* (5), Ant^{NS} (6), v^{36f} (9) and Bx^1 and Bx^3 (7). McGinnis and Beckendorf (8) have described deletions with end points in, or near, *roo* elements inserted near the *sgs*4 gene. These deletions are associated with *Notched* mutations.

References: (1) Scherer *et al.* (1982).
(2) Scherer *et al.* (1981).
(3) Meyerowitz and Hogness (1982).
(4) O'Hare *et al.* (1984).

(5) Swaroop *et al.* (1985).
(6) Scott *et al.* (1983).
(7) Mattox and Davidson (1984).
(8) McGinnis and Beckendorf (1983).
(9) Searles and Voelker (1985).



LTR of a B104/roo element.

```
   1  TGTTCACACATGAACACGAATATATTTAAAGACTTACAATTTTGGGCTCC   50
  51  GTTCATATCTTATGTAAATGAATCGAGAGCGATAAATTATATTTAGGATT  100
 101  TTGTTATCTAAGGCGACATGGGTGCATTGCTCAAAAACATGTAATTTAAG  150
 151  TGCACACTACATGAGTCAGTCACTTGAGATCGTTCCCCGCCTCCTAAAAA  200
 201  TAGTCCCTTAGTGGGAGACCACAGATAAGGTCCTCGCCGCTCAAGATAGG  250
 251  CAGATGTGCCCGAGCGTGGGACCTCGATAAGGCGGGGACTATTTACGTAG  300
 301  GCCTCTGCGTAGGCCATTTACTTTAAGATGCGATTCTCATGTCACCTATT  350
 351  TAAACCGAAGATATTTCCAAATAAAATCAGTTTCTTACAAAAACTCAACG  400
 401  AGTAAAGTCTTCTTATTTGGGATTTTACA  429
```

springer

Length: 8.8 kb (1).

Map: From the map published by Karlik and Fyrberg (1). There are no sites for the enzyme BamHI.

Terminal inverted repeat: 6/10 bp (1).

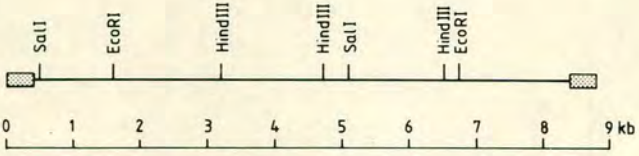
Target site duplication: 6 bp (1).

Approximate copy number: 6 (1).

Comments: First described by Karlik and Fyrberg (1) as an insertion within the gene for IFM-specific tropomyosin isoform. This is associated with the mutation *Ifm(3)3*. Only one copy of this element has been described in detail. This sequence was reported by Karlik and Fyrberg (1). The length of the target site duplication is uncertain. The 6 bp target site duplication deduced for this copy of the element is the sequence TATATA. Since this is the same when read on either strand, all, or part, of it could belong to the

inverted repeats at the ends of the LTRs, rather than being part of the target site.

Reference: (1) Karlik and Fyrberg (1985).



LTR of a springer element.

```
  1  AATTAATTAAATGTATGGTGCAGGTCCCTCGCCGCGGTCTCCGGCGTAGG   50
 51  TTGCAGGTAACGGGGGGTTCCTCTGTCACTGGGAGGCAGGGCGGTTGCCG  100
101  CAGACCTCTTCTCCTAGATTGGGAGATATGGTGGGAGAACGTCTCTCCGT  150
151  TGTTGACTGCCCTTAAGGCTAGCCAACCAATTCAATGATAACAGGCAGTT  200
201  AGCTGGAGTTAGATTGAAGGCGCGATGCGCTCTTTTATTGGAATACAAAT  250
251  CAAACTGACTATAAGCTACAAGGGAAAACATCATAGCGGCCTCTGCCAAT  300
301  GCGCAGAGCTTCTGCCGGCTATGCATGAGCTTCCGGCCAAATGCTTGGTC  350
351  AGCAATTTGACCGGTGCTGGTGTGCGGACGATCAGTCCGGTTAACTTAGT  400
401  TAACT  405
```

F

Length: variable, but consensus element is 4.7 kb (1).

Map: From the consensus map of an *F* element published by Di Nocera *et al.* (1). There are no sites for the enzymes ClaI, PvuI and XhoI.

Terminal inverted repeat: None (1).

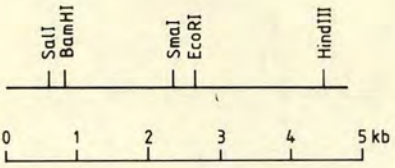
Target site duplication: Variable, *F* elements have been found associated with duplications of 8–13 bp (1).

Approximate copy number: 50 (1).

Comments: First described by Dawid *et al.* (2). *101F* was found inserted into a copy of the type I 28 S rDNA insertion sequence (2). This sequence was reported by Di Nocera *et al.* (1) for *101F*. An *F* element is associated with the mutation w^{i+A} (3). The w^1 mutation is associated with insertion of a 6 kb element having interrupted homology with *F* elements (4). The element '*Jiminy*', which was identified within the Bithorax complex by Bender *et al.* (5), is probably an *F* element.

References: (1) Di Nocera *et al.* (1983).
(2) Dawid *et al.* (1981).

(3) O'Hare *et al.* (1984).
(4) O'Hare *et al.* (1983).
(5) Bender *et al.* (1983*a*).



Ends of an F element

F101 left-hand end

```
   1  ATGAAGCATTTCGATCGCCGACGTGTGAAGACGTTTTTATCGAGCTCCGC   50
  51  ACAAAATCGGTTGTTTTGAGTGAAGTAAACGCCAAATAAAGTAAACTAAA  100
```

F101 right-hand end

```
 712  TACAACTGTTCAAAAAATAGTTAAGTTTTTAAACTTATTGTTAGTTCTTA  761
 762  TACAAGAAGATTCAATAAATAAAAGTAAAGTAAAAAAAAAAAAAAAAAG  811
```

FB

Length: variable (2).

Map: From the sequence of the FB4 element reported by Potter (1). There are no sites for the enzymes AvaI, BamHI, EcoRI, HpaI, PstI, SacI, SalI, SmaI and XhoI. The only HinfI and TaqI sites shown are those which lie within the inverted repeats.

Terminal inverted repeat: The length of the terminal inverted repeats of FB elements varies. The inverted repeats of FB4 are about 1 kb long (2).

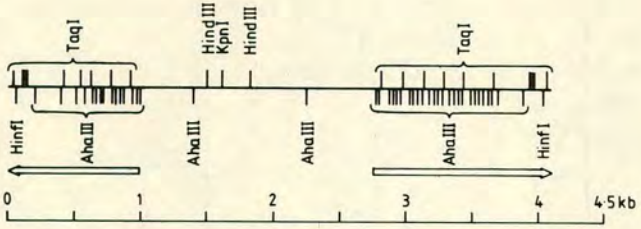
Target site duplication: 9 bp (1, 2, 3, 4).

Approximate copy number: 30 (2).

Comments: First described by Potter (7) as elements containing inverted repeat sequences. This sequence was reported by Potter (1). The putative transposable element HB1 lies between coordinates 1.1 and 2.75. An FB element is associated with the mutation w^{c} (5). FB elements flank DNA transposed by TE elements. A small TE element is associated with the mutation w^{DZL} (6). This is 13 kb long and has FB elements 2.2 and 3.8 kb long at its ends.

References: (1) Potter (1982*a*).
(2) Truett *et al.* (1981).
(3) Collins and Rubin (1982).
(4) O'Hare *et al.* (1984).

(5) Collins and Rubin (1982).
(6) Levis and Rubin (1982).
(7) Potter *et al.* (1980).



Ends of an FB element.

FB4 left-hand end

```
   1  AGCTCAAAGAAGCTGGGGTCGGAAAAATCGAATTTTTGAAATTTGAAAGC   50
  51  TGGAATCGTTTGCCCATTTTTTGCCCATGTTTGCCCACCAATTAGTTTTT  100
 101  TTTGCCCACGTCCAGTTTTTGAGATATGGATTTTCGAAAAAGTTCGAAAA  150
 151  TGTTCGAAAATCAAAAATTTCGCTTTTTTCAAATTTTTTTTTTTTTAAAT  200
```

FB4 right-hand end.

```
3890  ATTTAAAAAAAAAAAATTTTGAAAAAAGCGAAATTTTTGATTTTCGAAAA  3939
3940  TTTTCGAACTTTTTCGAAAATTCATATCTCAGAAACTGGACGTGGGCAAA  3989
3990  AAAAACTAATTGGTGGGCAAACATGGGCAAAAAATGGGCATACGATTCCA  4039
4040  GCTTTCAAATTTCAAAAATTCGATTTTTCCGACCCCAGCTTCTTTGAGCT  4089
```

G

Length: variable, the longest element so far is about 4 kb (1, 2).

Map: Taken from the map of the G1 element published by Di Nocera and Dawid (1). There are no sites for the enzyme SmaI.

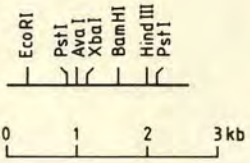
Target site duplication: 9 bp (2).

Approximate copy number: 10–20 (1, 2).

Comments: First described by Di Nocera and Dawid (1) as a sequence inserted within an F element. Three G elements, G1–3, were found interspersed with F sequences, probably the result of tandem duplication events. There is an oligo(A) sequence at the 3′ end of one strand of G1, and this at the right of the map above. The elements G2 and G3 are apparently identical to each other and lack the left-hand part of G1. They also have an extra 400 bp sequence, including an HindIII site, between the SalI/XhoI and BamHI sites. G elements have been found in tandem arrays inserted in the

non-transcribed spacer sequences of rDNA units. The chromosomal distribution of G elements is fairly stable by comparison with other transposable elements, as assayed by Southern transfer experiments, and they are concentrated in the chromocentric regions of polytene chromosomes (2). No polyA$^+$ transcripts complementary to G elements have been found in embryos, larvae, pupae, or adults (3).

References: (1) Di Nocera and Dawid (1983).
(2) Di Nocera *et al.* (1986).
(3) Dawid *et al.* (1981).



VIII. References

Ananiev, E. V., Gvozdev, V. A., Ilyin, Y. V., Tchurikov, N. A., and Georgiev, G. P. (1978). Reiterated genes with varying location in intercalary heterochromatin of *Drosophila melanogaster* polytene chromosomes. *Chromosoma* **70**, 1–17.

Anxolabehere, D., Kai, H., Nouaud, D., Periquet, G., and Ronsseray, S. (1984). The geographical distribution of P-M hybrid dysgenesis in *Drosophila melanogaster*. *Genet. Sel. Evol.* **16**, 15–26.

—— Nouaud, D., Periquet, G., and Tcher, P. (1985). P-element distribution in Eurasian populations of *Drosophila melanogaster*. A genetic and molecular analysis. *Proc. Natl. Acad. Sci. USA* **82**, 5418–5422.

Arkhipova, I. R., Gorelove, T. V., Ilyin, Y. V., and Schuppe, N. G. (1984). Reverse transcription of *Drosophila* mobile dispersed genetic element RNA: detection of intermediate forms. *Nucl. Acids Res.* **12**, 7533–7548.

Bayev, A. A., Krayev, A. S., Lyubomirskaya, N. V., Ilyin, Y. V., Skryabin, K. G., and Georgiev, G. P. (1980). The transposable element mdg3 in *Drosophila melanogaster* is flanked with perfect direct and mismatched inverted repeats. *Nucl. Acids Res.* **8**, 3263–3273.

—— Lyubomirskya, N. V., Dzhumagliev, E. B., Ananiev, E. V., Amiantova, I. G., and Ilyin, Y. V. (1984). Structural organisation of transposable element mdg4 from *Drosophila melanogaster* and nucleotide sequence of its long terminal repeats. *Nucl. Acids Res.* **12**, 3707–3723.

Bell, J. R., Bogardus, A. M., Schmidt, T., and Pellegrini, M. (1985). A new *copia*-like transposable element found in a Drosophila rDNA gene unit. *Nucl. Acids Res.* **13**, 3861–3871.

Bender, W., Akam, M., Korch, F., Beachy, P. A., Peifer, M., Spierer, P., Lewis, E. B., and Hognes, D. S. (1983*a*). Molecular genetics of the bithorax complex of *Drosophila melanogaster*. *Science* **221**, 23–29.

—— Spierer, P., and Hogness, D. S. (1983*b*). Chromosomal walking and jumping to isolate DNA from the *Ace* and *rosy* loci and the Bithorax Complex in *Drosophila melanogaster*. *J. Mol. Biol.* **168**, 17–33.

Bingham, P. M. (1980). The regulation of *white* locus expression: a dominant mutant allele at the *white* locus of *Drosophila melanogaster*. *Genetics* **95**, 341–353.

—— (1981). A novel dominant mutant allele at the *white* locus of *Drosophila melanogaster* is mutable. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 519–525.

—— and Judd, B. H. (1981). A copy of the *copia* transposable element is very tightly linked to the w^a allele at the *white* locus of *D. melanogaster*. *Cell* **25**, 705–711.

—— Kidwell, M. G., and Rubin, G. M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain specific transposon family. *Cell* **29**, 995–1004.

Boeke, J. D., Garfinkel, D. J., Styles, C. A., and Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500.

Bregliano, J. C. and Kidwell, M. G. (1983). Hybrid dysgenesis determinants. In *Mobile Genetic Elements* (J. Shapiro ed.) pp. 363–410. Academic Press, New York.

Brierley, H. L. and Potter, S. S. (1985). Distinct characteristics of loop sequences of two *Drosophila* foldback transposable elements. *Nucl. Acids Res.* **13**, 485–500.

Brookfield, J. F. Y., Montgomery, E., and Langley, C. H. (1984). Apparent absence of transposable elements related to the P elements of *D. melanogaster* in other species of *Drosophila*. *Nature* **310**, 330–332.

Bucheton, A., Paro, R., Sang, H. M., Pelisson, A., and Finnegan, D. J. (1984). The molecular basis of I-R hybrid dysgenesis: identification, cloning and properties of the I factor. *Cell* **38**, 153–163.

Cameron, J. R., Loh, E. Y., and Davis, R. W. (1979). Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* **16**, 739–751.

Campuzano, S., Carramolino, L., Cabrera, C. V., Ruiz-Gromez, M., Villares, R., Boromak, A., and Modelell, J. (1985). Molecular genetics of the *achaete–scute* gene complex of *D. melanogaster*. *Cell* **40**, 327–338.

—— Balcells, L., Villares, R., Carramolino, L., Garcia-Alonso, L., and Modelell, J. (1986). Excess function *Hairy-wing* mutations caused by gypsy and copia transposable elements inserted within structural genes of the *achaete–scute* locus of *Drosophila*. *Cell* **44**, 303–312.

Carbonara, B. D. and Gehring, W. J. (1985). Excision of *copia* element in a revertant of the *white–apricot* mutation of *Drosophila melanogaster* leaves behind one long terminal repeat. *Mol. Gen. Genet.* **199**, 1–6.

Chia, W., McGill, S., Karp, R., Gubb, D., and Ashburner, M. (1985). Spontaneous excision of a large composite transposable element of *Drosophila melanogaster*. *Nature* **316**, 81–83.

Clare, J. and Farabaugh, P. (1985). Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. *Proc. Natl. Acad. Sci. USA* **82**, 2829–2833.

Collins, M. and Rubin, G. M. (1982). Structure of the *Drosophila* mutable allele, *white–crimson*, and its *white–ivory* and wild-type derivatives. *Cell* **30**, 71–79.

—— —— (1983). High-frequency precise excision of the *Drosophila* foldback transposable element. *Nature* **303**, 259–260.

—— —— (1984). Structure of chromosomal rearrangements induced by an FB transposable element in *Drosophila*. *Nature* **308**, 323–327.

Copeland, N. G., Hutchinson, K. W., and Jenkins, N. A. (1983). Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. *Cell* **33**, 379–387.

Daniels, S. B., Strausbaugh, L. D., Ehrman, L., and Armstrong, R. (1984). Sequences homologous to P elements occur in *Drosophila paulistorum*. *Proc. Natl. Acad. Sci. USA* **81**, 6794–6797.

Dawid, I. B., Olong, E. O., Di Nocera, P. P., and Pardue, M. L. (1981). Ribosomal insertion-like elements in *Drosophila melanogaster* are interspersed with mobile sequences. *Cell* **25**, 399–408.

Di Nocera, P. P. and Dawid, I. B. (1983). Interdigitated arrangement of two oligo(A)-terminated DNA sequences in *Drosophila*. *Nucl. Acids Res.* **11**, 5475–5482.

—— Digan, M. E., and Dawid, I. (1983). A family of oligo-adenylated transposable sequences in *Drosophila melanogaster*. *J. Mol. Biol.* **168**, 715–727.

—— Graziani, F., and Lavorgna, G. (1986). Genomic and structural organisation of *Drosophila melanogaster* G elements. *Nucl. Acids Res.*, in press.

Donehower, L. A. and Varmus, H. E. (1984). A mutant murine leukemia virus with a single missense codon in *pol* is defective in a function affecting integration. *Proc. Natl. Acad. Sci. USA* **81**, 6461–6465.

Dunsmuir, P., Brorien, W. J., Simon, M. A., and Rubin, G. M. (1980). Insertion of the *Drosophila* transposable element copia generates a 5 base pair duplication. *Cell* **21**, 576–579.

Duyk, G., Leis, J., Longiarce, M., and Skalka, A. M. (1983). Selective cleavage in the avian retroviral long terminal repeat sequence by the endonuclease associated with the $\alpha\beta$ form of avian reverse transcriptase. *Proc. Natl. Acad. Sci. USA* **80**, 6745–6749.

Echalier, G. and Ohanessian, A. (1969). *In vitro* culture of *Drosophila melanogaster embryonic cells*. *In vitro* **6**, 162–172.

Emori, Y., Shiba, T., Kanaya, S., Inouye, S., Yuki, S., and Saigo, K. (1985). The nucleotide sequences of *copia* and *copia*-related RNA in *Drosophila* virus-like particles. *Nature* **315**, 773–776.

Engels, W. R. (1979). Extrachromosomal control of mutability in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **79**, 4011–4015.

—— (1984). A *trans*-acting product needed for P factor transposition in *Drosophila*. *Science* **226**, 1194–1196.

—— and Preston, C. R. (1984). Formation of chromosome rearrangements by P factors in Drosophila. *Genetics* **107**, 657–678.

Fabijanski, S. and Pelligrini, M. (1982). A *Drosophila* ribosomal protein gene is located near repeated sequences including rDNA sequences, *Nucl. Acids Res.* **10**, 5979–5991.

Falkenthal, S. and Lengyel, J. (1980). Structure, translation and metabolism of the cytoplasmic copia ribonucleic acid of *Drosophila melanogaster*. *Biochemistry* **19**, 5842–5850.

Fawcett, D. H., Lister, C. K., and Finnegan, D. J. (1986). The transposable element controlling IR hybrid dysgenesis is similar to mammalian LINES. In preparation.

Finnegan, D. J. (1985). Transposble elements in eukaryotes. *Int. Rev. Cytol.* **93**, 281–326.

—— Rubin, G. M., Young, M. W., and Hogness, D. S. (1978). Repeated gene families in *Drosophila melanogaster*. *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1053–1063.

Flavell, A. J. (1984). The involvement of reverse transcriptase in the generation of extrachromosomal copia mobile genetic elements. *Nature* **310**, 514–516.

—— and Ish Horowicz, D. (1981). Extrachromosomal circular copies of the eukaryotic transposable element *copia* in cultured *Drosophila* cells. *Nature* **292**, 591–574.

—— and Ish Horowicz, D. (1983). The origin of extrachromosomal circular copia elements. *Cell* **34**, 415–419.

—— Ruby, S. W., Toole, J. J., Roberts, B. E., and Rubin, G. M. (1980). Translation and developmental regulation of RNA encoded by the eukaryotic transposable element copia. *Proc. Natl. Acad. Sci. USA* **77**, 7107–7111.

—— Levis, R., Simon, M. A., and Rubin, G. M. (1981). The 5′ termini of RNAs encoded by transposable element copia. *Nucl. Acids Res.* **9**, 6279–6291.

Freund, R. and Meselson, M. (1984). Long terminal repeat nucleotide sequence and

specific insertion of the gypsy transposon. *Proc. Natl. Acad. Sci. USA* **81**, 4462–4464.

Gehring, W. J. and Paro, R. (1980). Isolation of a hybrid plasmid with homologous sequences to a transposing element of *Drosophila melanogaster*. *Cell* **19**, 897–904.

Georgiev, G. P. (1984). Mobile genetic elements in animal cells and their biological significance. *Eur. J. Biochem.* **145**, 203–220.

—— Ilyin, Y. V., Ryskov, A. R., Tchurikov, N. A., and Yenikolopov, G. N. (1977). Isolation of eukaryotic DNA fragments containing structural genes and the adjacent sequences. *Science* **195**, 394–397.

Gerasimova, T. I. (1983). Genetic instability at the cut locus of *Drosophila melanogaster induced by the MR-R12* chromosome. *Mol. Gen. Genet.* **184**, 544–547.

—— Mirzokhi, L. J., and Georgiev, G. P. (1984). Transposition bursts in genetically unstable *Drosophila melanogaster*. *Nature* **309**, 714–716.

Goldberg, M. L., Sheen, J., Gehring, W. J., and Green, M. M. (1983). Unequal crossing-over associated with asymmetrical synapsis between nomadic elements in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci. USA* **80**, 5017–5021.

Golubovsky, M. D., Ivanov, Y. N., and Green, M. M. (1977). Genetic instability in *Drosophila melanogaster*: putative multiple insertion mutants at the singed bristles locus. *Proc. Natl. Acad. Sci. USA* **74**, 2973–2975.

Green, M. M. (1959). Spatial and functional properties of pseudoalleles at the *white* locus in *Drosophila melanogaster*. *Heredity* **13**, 302–315.

—— (1967). The genetics of a mutable gene at the white locus of *Drosophila melanogaster*. *Genetics* **56**, 467–482.

—— (1977). Genetic instability in *Drosophila melanogaster*. *De novo* induction of putative insertion mutants. *Proc. Natl. Acad. Sci. USA* **74**, 3490–3493.

Gubb, D., Shelton, M., Roote, J., McGill, S., and Ashburner, M. (1984) The genetic analysis of a large transposing element of *Drosophila melanogaster*. *Chromosoma* **91**, 54–64.

—— Roote, J., McGill, S., Shelton, M., and Ashburner, M. (1985). Interactions between *white* genes carried by a large transposing element and the *zeste*^{1} allele in *Drosophila melanogaster*. *Genetics* **112**, 551–575.

Haynes, S. R., Toomey, T. P., Leinswand, L., and Jelineck, W. R. (1981). The chinese hamster *Alu*-like sequence: a conserved highly repetitious, interspersed deoxynucleic acid sequence in mammals has a structure suggestive of a transposable element. *Mol. Cell. Biol.* **1**, 573–583.

Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D., and Leder, P. (1982). Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. *Nature* **296**, 321–325.

Houck, C. M., Rinehart, F. P., and Schmid, C. W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* **132**, 289–306.

Ikenaga, H. and Saigo, K. (1982). Insertion of a movable genetic element, *297*, into the T-A-T-A box for the H3 histone in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **79**, 4143–4147.

Ilyin, Y. V., Tchurikov, N. A., Ananiev, E. V., Ryskov, A. P., Yenikolopov, G. N., Limborska, S. A., Maleeva, N. E., Gvozdev, V. A., and Georgiev, G. P. (1978). Studies on the DNA fragments of mammals and *Drosophila* containing structural genes and adjacent sequences. *Cold Spring Harbor Symp. Quant. Biol.* **42**, 959–969.

—— Chmeliauskaite, V. G., Ananiev, E. V., and Georgiev, G. P. (1980*a*). Isolation and characterisation of a new family of mobile dispersed genetic elements, mdg3 in *Drosophila melanogaster*. *Chromosoma* **81**, 27–53.

—— Chmeliauskaite, V. G., Ananiev, E. V., Lyubomirskya, N. V., Bayev, A. A., and

Georgiev, G. P. (1980*b*). Mobile dispersed genetic element MDG1 of *Drosophila melanogaster*: structural organisation. *Nucl. Acid Res.* **8**, 5333–5346.

—— Chmeliauskaite, V. G., and Georgiev, G. P. (1980*c*). Mobile disperse genetic element MDG1 of *Drosophila melanogaster*: transcription pattern. *Nucl. Acids Res.* **8**, 3439–3457.

—— Chmeliauskaite, V. G., Kulguskin, V. V., and Georgiev, G. P. (1980*d*). Double-stranded sequences in RNA of *Drosophila melanogaster*: relation to mobile dispersed genes. *Nucl. Acids Res.* **8**, 5347–5361.

—— Schuppe, N. G., Lyubomirskya, N. V., Gorelova, T. V., and Archipova, I. R. (1984). Circular copies of mobile dispersed genetic elements in cultured *Drosophila melanogaster* cells. *Nucl. Acids Res.* **12**, 7517–7531.

Inouye, S., Yuki, S., and Saigo, K. (1984). Sequence-specific insertion of the *Drosophila* transposable element *17.6*. *Nature* **310**, 332–333.

Ising, G. and Block, K. (1981). Derivation-dependent distribution of insertion sites for a *Drosophila* transposon. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 527–549.

—— and Block, K. (1984). A transposon as a cytological marker in *Drosophila melanogaster*. *Mol. Gen. Genet.* **196**, 6–16.

—— and Ramel, C. (1976). The behaviour of a transposable element in *Drosophila melanogaster*. In *The Genetics and Biology of Drosophila* (M. Ashburner and E. Novitski eds.) Vol. 1b pp. 947–954. Academic Press, London.

James, H. (1980). An X-ray crystallographic approach to enzyme structure and function. *Can. J. Biochem.* **58**, 251–271.

Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houk, C. M., Deininger, P. L., and Schmid, C. W. (1980). Ubiquitous interspersed, repeated sequences in mammalian genomes. *Proc. Natl. Acad. Sci. USA* **77**, 1398–1402.

Junakovic, N. and Ballario, P. (1984). Circular extrachromosomal *copia*-like transposable elements in *Drosophila* cultured cells. *Plamid* **11**, 109–115.

Karch, F., Weiffenbach, B., Bender, W., Duncan, I., Celniker, S., Crosby, M., and Lewis, P. B. (1985). The abdominal region of the Bithorax Complex. *Cell* **43**, 81–96.

Karess, R. E. and Rubin, G. M. (1984). Analysis of P transposable element functions in *Drosophila*. *Cell* **38**, 135–146.

Karlik, C. C. and Fyrberg, E. E. (1985) An insertion within a variably spiced *Drosophila* tropomyosin gene blocks accumulation of only one encoded form. *Cell* **441**, 57–66.

Kidwell, M. G. (1983). Intraspecific hybrid sterility. In *Genetics and Biology of Drosophila* (M. Ashburner, H. L. Carson and J. N. Thompson eds.) Vol. 3c, pp. 125–153. Academic Press, London.

Kramerov, D. A., Grigorgan, A. A., Ryskov, A. P., and Georgiev, G. P. (1979). Long double stranded sequences (ds RNA-B) of nuclear pre-mRNA consist of a few highly abundant classes of sequence: evidence from DNA cloning experiments. *Nucl. Acids Res.* **6**, 697–713.

Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A., and Georgiev, G. P. (1980). The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. *Nucl. Acids Res.* **8**, 1201–1215.

Kugimiya, W., Ikenaga, H., and Saigo, K. (1983). Close relationship between the long terminal repeats of avian leukosis–sarcoma virus and *copia*-like movable genetic elements of *Drosophila*. *Proc. Natl. Acad. Sci. USA* **80**, 3193–3197.

Kulguskin, V. V., Ilyin, Y. V., and Georgiev, G. P. (1981). Mobile dispersed genetic

element MDG1 of *Drosophila melanogaster*: nucleotide sequence of long terminal repeats. *Nucl. Acid Res.* **9**, 3451–3463.

Laski, F. A., Rio, D. C., and Rubin, G. M. (1986). Tissue specificity of *Drosophila* P elements transposition is regulated at the level of mRNA splicing. *Cell* **44**, 7–19.

Levis, R. and Rubin, G. M. (1982). The unstable w^{DZL} mutation of *Drosophila* is caused by a 13 kilobase insertion that is imprecisely excised in phenotypic revertants. *Cell* **30**, 543–550.

—— Dunsmuir, P., and Rubin, G. M. (1980). Terminal repeats of the *Drosophila* transposable element copia: nucleotide sequence and genomic organisation. *Cell* **21**, 581–588.

—— Bingham, P. M., and Rubin, G. M. (1982). Physical map of the white locus of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **79**, 564–568.

—— O'Hare, K., and Rubin, G. M. (1984). Effects of transposable element insertions on RNA encoded by the *white* gene of Drosophila. *Cell* **38**, 471–481.

Lewis, E. B. (1949). Drosophila Information Service **23**, 59–60.

Martin, G., Wiernasz, D., and Schedl, P. (1983). Evolution of *Drosophila* repetitive-dispersed DNA. *J. Mol. Evol.* **19**, 203–213.

Mattox, W. M. and Davidson, N. (1984). Isolation and characterisation of the *Beadex* locus of *Drosophila melanogaster*: a putative *cis*-acting negative regulatory element for the *heldup-a* gene. *Mol. Cell. Biol.* **4**, 1343–1353.

McGinnis, W. and Beckendorf, S. K. (1983). Association of a *Drosophila* transposable element of the *roo* family with chromosomal deletion breakpoints. *Nucl. Acids Res.* **11**, 737–751.

—— Shermoen, A. W., and Beckendorf, S. K. (1983). A transposable element inserted just 5′ to a Drosophila glue protein gene alters gene expression and chromatin structure. *Cell* **34**, 75–84.

Meyerowitz, E. M. and Hogness, D. S. (1982). Molecular organisation of a Drosphila puff site that responds to ecdysone. *Cell* **28**, 165–176.

Modelell, J., Bender, W., and Meselson, M. (1983). *Drosophila melanogaster* mutations suppressible by the suppressor of Hairy-wing are insertions of a 7.3 kb mobile element. *Proc. Natl. Acad. Sci. USA* **80**, 1678–1682.

Montgomery, E. A. and Langley, C. H. (1983). Transposable elements in Mendelian populations II. distribution of three copia-like elements in natural populations. *Genetics* **104**, 473–483.

Mossie, K. G., Young, M. W., and Varmus, H. E. (1985). Extrachromosomal DNA forms of *copia*-like transposable elements, F elements and middle repetitive DNA sequences in *Drosophila melanogaster*. *J. Mol. Biol.* **182**, 31–43.

Mount, S. M. and Rubin, G. M. (1985). Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. *Mol. Cell. Biol.* **5**, 1630–1638.

O'Hare, K. (1985). The mechanism of control of P element transposition in *Drosophila melanogaster*. *Trends in Genetics* **1**, 250–254.

—— and Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell* **34**, 25–35.

—— Levis, R., and Rubin, G. M. (1983). Transcription of the *white* locus in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **80**, 6917–6921.

—— Murphy, C., Levis, R., and Rubin, G. M. (1984). DNA sequence of the *white* locus of *Drosophila melanogaster*. *J. Mol. Biol.* **180**, 437–455.

Panganiban, A. T. (1985). Retroviral DNA integration. *Cell* **42**, 5–6.

—— and Temin, H. M. (1984*a*). Circles with two tandem LTRs are precursors to integrated retroviral DNA. *Cell* **36**, 673–679.

—— and Temin, H. M. (1984*b*). The retrovirus *pol* gene encodes a product required

for DNA integration: identification of a retroviral *int* locus. *Proc. Natl. Acad. Sci. USA* **81**, 7885–7889.

Pardue, M. L. and Dawid, I. B. (1981). Chromosomal locations of two DNA segments that flank ribosomal insertion-like sequences in *Drosophila*: flanking sequences are mobile elements. *Chromosoma* **83**, 29–43.

Parkhurst, S. M. and Corces, V. G. (1985). *forked*, gypsys and suppressors in *Drosophila*. *Cell* **41**, 429–437.

Paro, R., Goldberg, M. L., and Gehring, W. J. (1983). Molecular analysis of large transposable elements carrying the *white* locus of *Drosophila melanogaster*. *EMBO J.* **2**, 853–860.

Patarca, R. and Heseltine, W. A. (1984). Letter. *Nature* **309**, 288.

Pelisson, A. (1981). The I-R system of hybrid dysgenesis in *Drosophila melanogaster*: are I factor insertions responsible for the mutator effect of the I-R interaction? *Mol. Gen. Genet.* **183**, 123–129.

Perronet, F., Rollet, E., Becker, J. L., Maisonhaute, C., Echalier, G., and Belpomme, M. (1986). From transcript modulations to protein phosphorylation: a short survey of some ecdysteroid effects. *Insect Biochem.* **16**, in press.

Pirotta, V. and Brockl, C. (1984). Transcription of the *Drosophila white* locus and some of its mutants. *EMBO J.* **3**, 563–568.

Potter, S. S. (1982*a*). DNA sequence of a foldback transposable element in *Drosophila*. *Nature* **297**, 201–204.

—— (1982b). DNA sequence analysis of a *Drosophila* foldback transposable element rearrangement. *Mol. Gen. Genet.* **188**, 107–110.

—— Brorien, W. J., Dunsmuir, P., and Rubin, G. M. (1979). Transposition of elements of the 412, copia and 297 gene families in *Drosophila*. *Cell* **17**, 415–427.

—— Truett, M., Phillips, M., and Maher, A. (1980) Eukaryotic transposable genetic elements with inverted terminal repeats. *Cell* **20**, 639–647.

Roeder, G. S. and Fink, G. R. (1983). Transposable elements in yeast. In *Mobile Genetic Elements* (J. A. Shapiro ed.) pp. 299–328. Academic Press, New York.

—— Beard, C., Smith, M., and Keranen, S. (1985). Isolation and characterisation of the *SPT2* gene. A negative regulator of *Ty*-controlled yeast gene expression. *Mol. Cell. Biol.* **5**, 1543–1553.

Rogers, J. E. (1983). Retroposons defined. *Nature* **301**, 460.

—— (1985). The origin and evolution of retroposons. *Int. Rev. Cytol.* **93**, 188–280.

Rubin, G. M. (1983). Dispersed repetitive DNAs in *Drosophila*. In *Mobile Genetic Elements* (J. A. Shapiro ed.), pp. 329–361. Academic Press, New York.

—— and Spradling, A. C. (1982). Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218**, 348–353.

—— Finnegan, D. J., and Hogness, D. S. (1976). The chromosomal arrangement of coding sequences in a family of repeated genes. In *Progress in Nucleic Acid Research* (W. Cohn and E. Volkin ed.) Vol. 19, pp. 221–226. Academic Press, New York.

—— Brorien, W. J., Dunsmuir, P., Flavell, A. J., Strobel, J. J., Toole, J. J. and Young, E. (1981). '*copia*-like' transposable elements in the *Drosophila* genome. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 619–628.

—— Kidwell, M. G., and Bingham, P. M. (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell* **29**, 987–994.

Saigo, K., Millstein, L., and Thomas, C. A. (1981). The organisation of *Drosophila melanogaster* histone genes. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 815–827.

—— Kugimya, W., Matsuo, Y., Inouye, S., Yoshioka, K., and Yuki, S. (1984). Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. *Nature* **312**, 659–661.

Sang, H. M., Pelisson, A., Bucheton, A., and Finnegan, D. J. (1984). Molecular lesions associated with *white* gene mutations induced by I-R hybrid dysgenesis in *Drosophila melanogaster*. *EMBO J.* **3**, 3079–3085.

Scherer, G., Telford, J., Baldari, C., and Pirrotta, V. (1981). Isolation of cloned genes differentially expressed at early and late stages of *Drosophila* embryonic development. *Developmental Biol.* **86**, 438–447.

—— Tschudi, C., Perera, J., Delius, H., and Pirrotta, V. (1982). *B104*, a new dispersed repeated gene family in *Drosophila melanogaster* and its analogies with retroviruses. *J. Mol. Biol.* **157**, 435–452.

Schmid, C. W., Manning, J. E., and Davidson, N. (1975). Inverted repeat sequences in the *Drosophila* genome. *Cell* **5**, 159–172.

Schwartz, D. E., Tizard, R., and Gilbert, W. (1983). Nucleotide sequence of Rous sarcoma virus. *Cell* **32**, 853–869.

Schwartz, H. E., Lockett, T. J., and Young, M. W. (1982). Analysis of transcripts from two families of nomadic DNA. *J. Mol. Biol.* **157**, 49–58.

Schwartzberg, P., Colicelli, J., and Goff, S. P. (1984). Construction and analysis of deletion mutants in the *pol* gene of Moloney murine leukemia virus: a new viral function required for productive infection. *Cell* **37**, 1043–1052.

Scott, M. L., McKereghan, K., Kaplan, H. S., and Fry, K. E. (1981). Molecular cloning and partial characterisation of unintegrated linear DNA from gibbon ape leukemia virus. *Proc. Natl. Acad. Sci. USA* **78**, 4213–4217.

Scott, M. P., Weiner, A. J., Hazelrigg, T. F., Polisky, B. A., Pirrotta, V., Scalenghe, F., and Kaufman, T. C. (1983). The molecular organization of the *Antennapedia* locus of *Drosophila*. *Cell* **35**, 763–776.

Searles, L. L. and Voelker, R. A. (1986) Molecular characterisation of the *Drosophila vermillion* locus and its suppressible alleles. *Proc. Natl. Acad. Sci. USA* **83**, 404–408.

Seiki, M., Hattori, S., Hirayami, Y., and Yoshida, M. (1983). Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc. Natl. Acad. Sci. USA* **80**, 3618–3622.

Shapiro, J. A. (1983). *Mobile Genetic Elements*. Academic Press, New York.

Shepherd, B. M. and Finnegan, D. J. (1984). Structure of circular copies of the 412 transposable element present in *Drosophila melanogaster* tissue culture cells, and isolation of a free 412 long terminal repeat. *J. Mol. Biol.* **180**, 21–40.

Shiba, T. and Saigo, K. (1983). Retrovirus-like particles containing RNA homologous to the transposable element *copia* in *Drosophila melanogaster*. *Nature* **302**, 119–124.

Shinnick, T. M., Lerner, R. A., and Sutcliffe, J. G. (1981). Nucleotide sequence of Moloney murine leukemia virus. *Nature* **293**, 543–548.

Simmons, M. J. and Lim, J. K. (1980). Site-specificity of mutations arising in dysgenic hybrids of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **77**, 6042–6046.

Singer, M. F. and Skowronski, J. (1985). Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. *Trends in Biochemical Sciences* **10**, 119–122.

Snyder, M. P., Kimbrell, D., Hunkapiller, M., Hill, R., Fristrom, J., and Davidson, N. (1982). A transposable element that splits the promoter region inactivates a *Drosophila* cuticle protein gene. *Proc. Natl. Acad. Sci. USA* **79**, 7430–7434.

Spradling, A. C. and Rubin, G. M. (1981). *Drosophila* genome organisation: conserved and dynamic aspects. *Ann. Rev. Genet.* **15**, 219–264.

—— (1982). Transposition of cloned P elements into *Drosophila* germ line chromosomes. *Science* **218**, 341–347.

Strobel, E., Dunsmuir, P., and Rubin, G. M. (1979). Polymorphisms in the chromo-

somal locations of elements of the 412, copia and 297 dispersed repeated gene families in *Drosophila*. *Cell* **17**, 429–439.

Swaroop, A., Paco-Larsen, M. E., and Garen, A. (1985). Molecular genetics of a transposon-induced dominant mutation in the *Drosophila* locus Glued. *Proc. Natl. Acad. Sci. USA* **82**, 1751–1755.

Sweet, H. O. (1983). Dilute suppressor, a new suppressor gene in the house mouse. *J. Hered.* **74**, 305–306.

Tchurikov, N. A., Ilyin, Y. V., Ananiev, E. V., and Georgiev, G. P. (1978). The properties of gene Dm225 a representative of dispersed repetitive genes in *Drosophila melanogaster*. *Nucl. Acids Res.* **6**, 2169–2187.

—— Zelentosova, E. S., and Georgiev, G. P. (1980). Clusters containing different dispersed genes in the genome of *Drosophila melanogaster*. *Nucl. Acids Res.* **8**, 1243–1258.

—— Ilyin, Y. V., Skyrabin, K. G., Ananiev, A. S., Krayev, A. S., Zelentsova, E. S., Kulguskin, V. V., Lyubomirskaya, N. V., and Georgiev, G. P. (1981). General properties of mobile dispersed genetic elements in *Drosophila melanogaster*. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 655–665.

Truett, M. A., Jones, R. S., and Potter, S. S. (1981). Unusual structure of the FB family of transposable elements in *Drosophila*. *Cell* **24**, 753–763.

van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T., and Gesteland, R. F. (1981). Direct repeats flank three small nuclear RNA pseudogenes in the human genome. *Cell* **26**, 11–17.

Varmus, H. E. (1983). Retroviruses. In *Mobile Genetic Elements* (J. A. Shapiro ed.) pp. 411–505. Academic Press, New York.

Voelker, R. A., Greenleaf, A. L., Gyurkovics, H., Wisely, G. B., Huang, S., and Searles, I. I. (1984). Frequent imprecise excision among reversions of a *P* element-caused lethal mutation in *Drosophila*. *Genetics* **107**, 279–294.

Weiss, R., Teich, N., and Varmus, H. (1984). *RNA Tumor Viruses Part I.* Cold Spring Harbor Laboratory, Cold Spring Harbor.

—— (1985). *RNA Tumor Viruses Part II.* Cold Spring Harbor Laboratory, Cold Spring Harbor.

Wensink, P. W., Tabata, S., and Pachi, C. (1979). The clustered and scrambled arrangement of moderately repetitive elements in *Drosophila* DNA. *Cell* **18**, 1231–1246.

Wilde, C. D., Crowther, C. E., Cripe, T. P., Lee, M. G., and Cowans, N. J. (1982). Evidence that a human β tubulin pseudogene is derived from its corresponding mRNA, *Nature* **297**, 83–84.

Will, B. M., Bayev, A. A., and Finnegan, D. J. (1981). Nucleotide sequence of terminal repeats of 412 transposable elements of *Drosophila melanogaster*. *J. Mol. Biol.* **153**, 897–915.

Young, M. W. (1979). Middle repetitive DNA: a fluid component of the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **76**, 6274–6278.

—— and Schwartz, H. E. (1981). Nomadic gene families in *Drosophila*. *Cold Spring Harbor Symp. Quant. Biol.* **45**, 629–640.

Zachar, Z. and Bingham, P. M. (1982). Regulation of *white* locus expression: the structure of mutant alleles at the white locus of *Drosophila melanogaster*. *Cell* **30**, 529–541.